Ann Oper Res (2013) 208:63–94
DOI 10.1007/s10479-011-1047-4

Derman's book as inspiration: some results on LP for MDPs

Lodewijk Kallenberg

Published online: 4 January 2012

Abstract In 1976 I was looking for a suitable subject for my PhD thesis. My thesis advisor Arie Hordijk and I found a lot of inspiration in Derman's book (Finite state Markovian decision processes, Academic Press, New York, 1970). Since that time I was interested in linear programming methods for Markov decision processes. In this article I will describe some results in this area on the following topics: (1) MDPs with the average reward criterion; (2) additional constraints; (3) applications. These topics are the main elements of Derman's book.

1 Introduction

When Arie Hordijk was appointed at the Leiden University in 1976, I became his first PhD student in Leiden. Hordijk was the successor of Guus Zoutendijk, who has chosen to leave the university for a position as chairman of the executive board of the Delta Lloyd Group. Zoutendijk was the supervisor of my master thesis and a leading expert in linear and non-linear programming. Looking for a PhD project Hordijk suggested linear programming (for short, LP) for the solution of Markov Decision Processes (for short, MDPs). LP for MDPs was introduced by D'Epenoux (1960) for the discounted case. De Ghellinck (1960) as well as Manne (1960) obtained LP formulations for the average reward criterion in the irreducible case. The first analysis of LP for the multichain case was given by Denardo and Fox (1968). Our interest was raised by Derman's remark (Derman 1970, p. 84): "No satisfactory treatment of the dual program for the multiple class case has been published".

We started to work on this subject and succeeded to present a satisfactory treatment of the dual program for multichained MDPs. We proved a theorem from which a simple algorithm follows for the determination of an optimal deterministic policy (Hordijk and Kallenberg 1979). In Sect. 3 we describe this approach. Furthermore, we present in Sect. 3 some examples which show the essential difference between irreducible, unichained and multichained MDPs. These examples show for general MDPs:

L. Kallenberg (✉)
Mathematical Institute, University of Leiden, P.O. Box 9512, 2300 RA Leiden, The Netherlands
e-mail: kallenberg@math.leidenuniv.nl

1. An extreme optimal solution of the dual program may have in some state more than one positive variable and consequently an extreme feasible solution of the dual program may correspond to a nondeterministic policy (Example 2).
2. Two different solutions may correspond to the same deterministic policy (Example 3).
3. An nonoptimal solution of the dual program may correspond to an optimal deterministic policy (Example 4).
4. The results of the unichain case cannot be generalized to the general single chain case (Example 5).

The second topic of this article concerns additional constraints. Chapter 7 of Derman's book deals with this subject and has as title "State-action frequencies and problems with constraints". This chapter may be considered as the starting point for the study of MDPs with additional constraints.

For unichained MDPs with additional constraints, Derman has shown that an optimal policy can be found in the class of stationary policies. We have generalized these results in the sense that for multichained MDPs stationary policies are not sufficient; however, in that case there exists an optimal policy in the class of Markov policies. This subject is presented in Sect. 4.

Derman's book also deals with some applications, for instance optimal stopping and replacement problems. In the last part, Sect. 5, of this paper we will discuss LP methods for the following applications:

1. Optimal stopping problems.
2. Replacement problems:
   (a) General replacement problems;
   (b) Replacement problems with increasing deterioration;
   (c) Skip to the right problems with failure;
   (d) Separable replacement problems.
3. Multi-armed bandit problems.
4. Separable problems with both the discounted and the average reward criterion.

2 Notations and definitions

Let S be the finite *state space* and $A(i)$ the finite *action set* in state $i \in S$. If in state i action $a \in A(i)$ is chosen, then a *reward* $r_i(a)$ is earned and $p_{ij}(a)$ is the *transition probability* that the next state is state j.

A *policy* R is a sequence of decision rules: $R = (\pi^1, \pi^2, \ldots, \pi^t, \ldots)$, where π^t is the decision rule at time point t, $t = 1, 2, \ldots$. The *decision rule* π^t at time point t may depend on all available information on the system until time t, i.e., on the states at the time points $1, 2, \ldots, t$ and the actions at the time points $1, 2, \ldots, t-1$.

Let C denote the set of all policies. A policy is said to be *memoryless* if the decision rules π^t are independent of the history; it depends only on the state at time t. We call $C(M)$ the set of the memoryless policies. Memoryless policies are also called *Markov policies*.

If a policy is memoryless and the decision rules are independent of the time point t, then the policy is called *stationary*. Hence, a stationary policy is determined by a nonnegative function π on $S \times A$, where $S \times A = \{(i, a) \mid i \in S,\ a \in A(i)\}$, such that $\sum_a \pi_{ia} = 1$ for every $i \in S$. The stationary policy $R = (\pi, \pi, \ldots)$ is denoted by π^∞. The set of stationary policies is notated by $C(S)$.

If the decision rule π of a stationary policy is nonrandomized, i.e., for every $i \in S$, we have $\pi_{ia} = 1$ for exactly one action a, then the policy is called *deterministic*. A deterministic

policy can be described by a function f on S, where $f(i)$ is the chosen action in state i. A deterministic policy is denoted by f^∞ and the set of deterministic policies by $C(D)$.

A matrix $P = (p_{ij})$ is a *transition matrix* if $p_{ij} \geq 0$ for all (i, j) and $\sum_j p_{ij} = 1$ for all i. Notice that P is a *stationary Markov chain*. For a Markov policy $R = (\pi^1, \pi^2, \dots)$ the *transition matrix* $P(\pi^t)$ is defined by

$$\{P(\pi^t)\}_{ij} = \sum_a p_{ij}(a)\pi^t_{ia}$$

and the vector $r(\pi^t)$, defined by

$$\{r(\pi^t)\}_i = \sum_a r_i(a)\pi^t_{ia},$$

is called the *reward vector*.

Let the random variables X_t and Y_t denote the state and action at time t. Given starting state i, policy R and a *discount factor* $\alpha \in (0, 1)$, the *discounted reward* and the *average reward* are denoted by $v^\alpha_i(R)$ and $\phi_i(R)$, respectively, and defined by

$$v^\alpha_i(R) = \sum_{t=1}^{\infty} \alpha^{t-1} \mathbb{E}_{i,R}\{r_{X_t}(Y_t)\}$$

and

$$\phi_i(R) = \liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{i,R}\{r_{X_t}(Y_t)\},$$

respectively.

The *value vectors* v^α and ϕ for discounted and average rewards are defined by $v^\alpha_i = \sup_R v^\alpha_i(R)$, $i \in S$, and $\phi_i = \sup_R \phi_i(R)$, $i \in S$, respectively.

A policy R^* is a *discounted optimal policy* if $v^\alpha_i(R^*) = v^\alpha_i$, $i \in S$; similarly, R^* is an *average optimal policy* if $\phi_i(R^*) = \phi_i$, $i \in S$. It is well known that, for both discounted as average rewards, an optimal policy exists and can be found within $C(D)$, the class of deterministic policies.

An MDP is called *irreducible* if, for all deterministic decision rules f, in the Markov chain $P(f)$ all states belong to a single ergodic class.

An MDP is called *unichained* if, for all deterministic decision rules f, in the Markov chain $P(f)$ all states belong to a single ergodic class plus a (perhaps empty and decision rule dependent) set of transient states. In the *weak unichain case* every optimal deterministic policy f^∞ has a unichain Markov chain $P(f)$; in the *general single chain case* at least one optimal deterministic policy f^∞ has a unichain Markov chain $P(f)$;

An MDP is called *multichained* if there may be several ergodic classes and some transient states; these classes may vary from policy to policy.

An MDP is *communicating* if for every $i, j \in S$ there exists a deterministic policy f^∞, which may depend on i and j, such that in the Markov chain $P(f)$ state j is accessible from state i.

It is well known that for irreducible, unichained and communicating MDPs the value vector has identical components. Hence, in these cases one uses, instead of a vector, a scalar ϕ for the value.

3 LP for MDPs with the average reward criterion

3.1 The irreducible case

In Chap. 6, pp. 78–80, of Derman's book the following result can be found, which originates from Manne (1960).

Theorem 1 *Let* (v^*, u^*) *and* x^* *be optimal solutions of* (1) *and* (2), *respectively, where*

$$\min\left\{v \mid v + \sum_j \{\delta_{ij} - p_{ij}(a)\}u_j \geq r_i(a),\ (i,a) \in S \times A\right\} \tag{1}$$

and

$$\max\left\{\sum_{i,a} r_i(a)x_i(a) \,\middle|\, \begin{array}{ll} \sum_{i,a}\{\delta_{ij} - p_{ij}(a)\}x_i(a) = 0, & j \in S \\ \sum_{i,a} x_i(a) = 1 & \\ x_i(a) \geq 0, & i \in S, a \in A(i) \end{array}\right\}. \tag{2}$$

Let f_*^∞ *be such that* $x_i^*(f_*(i)) > 0$, $i \in S$. *Then,* f_*^∞ *is well defined and an average optimal policy. Furthermore,* $v^* = \phi$, *the value.*

3.2 The unichain case

Theorem 2 *Let* (v^*, u^*) *and* x^* *be optimal solutions of* (1) *and* (2), *respectively. Let* $S_* = \{i \mid \sum_a x_i^*(a) > 0\}$. *Choose* f_*^∞ *such that* $x_i^*(f_*(i)) > 0$ *if* $i \in S_*$ *and choose* $f_*(i)$ *arbitrarily if* $i \notin S_*$. *Then,* f_*^∞ *is an average optimal policy. Furthermore,* $v^* = \phi$, *the value.*

This linear programming result for unichained MDPs was derived by Denardo (1970). I suppose that Derman was also aware of this result, although it was not explicitly mentioned in his book. Theorem 2 on p. 75 and the subsequent text on p. 76 are the reason of my supposition. The result of Theorem 2, but with a different proof, is part of my thesis (Kallenberg 1980), which was also published in Kallenberg (1983).

3.3 The communicating case

Since the value vector ϕ is constant in communicating MDPs, the value ϕ is the unique v^*-part of an optimal solution (v^*, u^*) of the linear program (1). One would expect that an optimal policy could also be obtained from the dual program (2). The next example shows that—in contrast with the irreducible and the unichain case—in the communicating case the optimal solution of the dual program doesn't provide an optimal policy, in general.

Example 1 $S = \{1,2,3\}$; $A(1) = \{1,2\}$, $A(2) = \{1,2,3\}$, $A(3) = \{1,2\}$. $r_1(1) = 0$, $r_1(2) = 2$; $r_2(1) = 1$, $r_2(2) = 1$, $r_2(3) = 3$; $r_3(1) = 2$; $r_3(2) = 4$. $p_{12}(1) = p_{11}(2) = p_{23}(1) = p_{21}(2) = p_{22}(3) = p_{32}(1) = p_{33}(2) = 1$ (other transitions are 0). This is a multichain and communicating model. The value is 4 and f_*^∞ with $f_*(1) = f_*(2) = 1$, $f_*(3) = 2$ is the unique optimal deterministic policy.

The primal linear program (1) becomes for this model

$$\min\left\{v \,\middle|\, \begin{array}{l} v + u_1 - u_2 \geq 0;\ v \geq 2;\ v + u_2 - u_3 \geq 1;\ v - u_1 + u_2 \geq 1 \\ v \geq 3;\ v - u_2 + u_3 \geq 2;\ v \geq 4 \end{array}\right\}$$

with optimal solution $v^* = 4$; $u_1^* = 0$, $u_2^* = 3$, $u_3^* = 5$ (v^* is unique; u^* is not unique).

The dual linear program is

$$\text{maximize } 2x_1(2) + x_2(1) + x_2(2) + 3x_2(3) + 2x_3(1) + 4x_3(2)$$

subject to

$$\begin{aligned}
x_1(1) \qquad\qquad - x_2(2) \qquad\qquad\qquad\qquad &= 0\\
-x_1(1) \qquad x_2(1) + x_2(2) \qquad - x_3(1) \qquad &= 0\\
\qquad\qquad - x_2(1) \qquad\qquad + x_3(1) \qquad &= 0\\
x_1(1) + x_1(2) + x_2(1) + x_2(2) + x_2(3) + x_3(1) + x_3(2) &= 1\\
x_1(1),\ x_1(2),\ x_2(1),\ x_2(2),\ x_2(3),\ x_3(1),\ x_3(2) &\geq 0
\end{aligned}$$

For the optimal solution x^*, we obtain: $x_1^*(1) = x_1^*(2) = x_2^*(1) = x_2^*(2) = x_2^*(3) = x_3^*(1) = 0$; $x_3^*(2) = 1$ (this solution is unique).

Proceeding as if this were a unichain model, we choose arbitrary actions in the states 1 and 2. Clearly, this approach may generate a nonoptimal policy.

So, we are not able—in general—to derive an optimal policy from the dual program (2). However, it is possible to find an optimal policy with some additional work. In Example 1 we have seen that the optimal solution x^* provides an optimal action in state 3, which is the only state of $S_* = \{i \mid \sum_a x_i^*(a) > 0\}$. The next theorem shows that the states of S_* always provide optimal actions. For the proof we refer to Kallenberg (2010).

Theorem 3 *Let x^* be an extreme optimal solution of (2). Take any policy f_*^∞ such that $x_i^*(f_*(i)) > 0$, $i \in S_*$. Then, $\phi_j(f_*^\infty) = \phi$, $j \in S_*$.*

Note that $S_* \neq \emptyset$ (because $\sum_{i,a} x_i^*(a) = 1$) and that we can find, by Theorem 3, optimal actions $f_*(i)$ for all $i \in S_*$. Furthermore, one can easily show that S_* is closed in the Markov chain $P(f_*)$.

Since we have a communicating MDP, one can find for each $i \notin S_*$ an action $f_*(i)$ such that in the Markov chain $P(f_*)$ the set S_* is reached from state i with a strictly positive probability after one or more transitions. So, the set $S \backslash S_*$ is transient in the Markov chain $P(f_*)$. Therefore, the following search procedure provides the remaining optimal actions for the states $S \backslash S_*$.

Search procedure

1. If $S_* = S$: stop;
   Otherwise go to step 2.
2. Pick a triple (i, a, j) with $i \in S \backslash S_*$, $a \in A(i)$, $j \in S_*$ and $p_{ij}(a) > 0$.
3. $f_*(i) := a$, $S_* := S_* \cup \{i\}$ and go to step 1.

A second way to find an optimal policy for communicating MDPs is based on the following theorem which is due to Filar and Schultz (1988).

Theorem 4 *An MDP is communicating if and only if for every $b \in \mathbb{R}^{|S|}$ such that $\sum_j b_j = 0$ there exists a $y \in \mathbb{R}_+^{|S \times A|}$ such that $\sum_{i,a} \{\delta_{ij} - p_{ij}(a)\} y_i(a) = b_j$ for all $j \in S$.*

The following procedure also yields an optimal deterministic policy. This is based on results for multichained MDPs which are discussed in Sect. 3.4.

Determination y-variables

1. Choose $\beta \in \mathbb{R}^{|S|}$ such that $\beta_j > 0$, $j \in S$ and $\sum_j \beta_j = 1$.
2. Let $b_j = \beta_j - \sum_a x_j^*(a)$, $j \in S$.
3. Determine $y^* \in \mathbb{R}_+^{|S \times A|}$ such that $\sum_{i,a}\{\delta_{ij} - p_{ij}(a)\}y_i^*(a) = b_j$, $j \in S$.
4. Choose $f_*(i)$ such that $y_i^*(f_*(i)) > 0$ for all $i \in S \backslash S_*$.

Example 1 (continued)

Search procedure:
$S_* = \{3\}$.
$i = 2$; $a = 1$; $j = 3$; $f_*(2) = 1$; $S_* = \{2, 3\}$.
$i = 1$; $a = 1$; $j = 2$; $f_*(1) = 1$; $S_* = \{1, 2, 3\}$.
Determination y-variables:
Choose $\beta_1 = \beta_2 = \beta_3 = \frac{1}{3}$.
Let $b_1 = \frac{1}{3}$, $b_2 = \frac{1}{3}$, $b_3 = -\frac{2}{3}$.

The system $\sum_{i,a}\{\delta_{ij} - p_{ij}(a)\}y_i(a) = b_j$, $j \in S$ becomes:

$$\begin{aligned} y_1(1) \qquad\quad - y_2(2) \qquad\quad &= \tfrac{1}{3} \\ -y_1(1) + y_2(1) + y_2(2) - y_3(1) &= \tfrac{1}{3} \\ -y_2(1) \qquad\quad + y_3(1) &= -\tfrac{2}{3} \end{aligned}$$

with a nonnegative solution $y_1^*(1) = \frac{1}{3}$, $y_2^*(1) = \frac{2}{3}$, $y_2^*(2) = y_3^*(1) = 0$ (this solution is not unique). Choose $f_*(1) = f_*(2) = 1$.

Remarks

1. The verification of an irreducible or communicating MDP is computationally easy (see Kallenberg 2002); generally, the verification of a unichain MDP is $\mathcal{NP}$-complete as shown by Tsitsiklis (2007).
2. It turns out that the approach with the search procedure can also be used for the weak unichain case.

3.4 The multichain case

For multichained MDPs the programs (1) and (2) are not sufficient. For general MDPs the following dual pair of linear programs were proposed by Denardo and Fox (1968):

$$\min \left\{ \sum_j \beta_j v_j \;\middle|\; \begin{array}{ll} \sum_j \{\delta_{ij} - p_{ij}(a)\}v_j \geq 0, & (i,a) \in S \times A \\ v_i + \sum_j \{\delta_{ij} - p_{ij}(a)\}u_j \geq r_i(a), & (i,a) \in S \times A \end{array} \right\} \tag{3}$$

and

$$\max \left\{ \sum_{(i,a)} r_i(a)x_i(a) \;\middle|\; \begin{array}{ll} \sum_{i,a}\{\delta_{ij} - p_{ij}(a)\}x_i(a) = 0, & j \in S \\ \sum_a x_j(a) + \sum_{i,a}\{\delta_{ij} - p_{ij}(a)\}y_i(a) = \beta_j, & j \in S \\ x_i(a), y_i(a) \geq 0, \quad i \in S, a \in A(i) & \end{array} \right\}, \tag{4}$$

where $\beta_j > 0$ for all $j \in S$.

In Denardo and Fox (1968) it was shown that if (v^*, u^*) is an optimal solution of the primal problem (3), than $v^* = \phi$, the value vector.

Notice that if the value vector ϕ is constant, i.e., ϕ has identical components, then $\sum_j \{\delta_{ij} - p_{ij}(a)\} v_j^* = \sum_j \{\delta_{ij} - p_{ij}(a)\}\phi = \{1-1\}\phi = 0$. Hence, the first set of inequalities of (3) is superfluous and (3) can be simplified to (1) with as dual program (2).

Furthermore, Denardo and Fox have derived the following result (see pp. 73–75 in Derman 1970).

Lemma 1 *Let $f_*^\infty \in C(D)$ be an optimal policy and let $(v^* = \phi, u^*)$ be an optimal solution of the primal program (3). Then,*

$$\begin{cases} \sum_j \{\delta_{ij} - p_{ij}(f_*)\}\phi_j = 0, & i \in S \\ \phi_i + \sum_j \{\delta_{ij} - p_{ij}(f_*)\} u_j^* = r_i(f_*), & i \in R(f_*) \end{cases}$$

where $R(f_*) = \{i \mid i$ is recurrent in the Markov chain $P(f_*)\}$.

Lemma 1 asserts that in any optimal solution of the primal program (3) one can always select actions $f_*(i)$ such that $\sum_j \{\delta_{ij} - p_{ij}(f_*)\}\phi_j = 0$, $i \in S$, and $\phi_i + \sum_j \{\delta_{ij} - p_{ij}(f_*)\} u_j^* = r_i(f_*)$ for all i in a nonempty subset $S(f_*)$ of S. Furthermore, the following result holds, given such policy f_*^∞ and a companion $S(f_*)$ (see pp. 75–76 in Derman 1970).

Lemma 2 *If all states of $S \backslash S(f_*)$ are transient in the Markov chain $P(f_*)$, then policy f_*^∞ is an average optimal policy.*

If we are fortunate in our selection of f_*^∞, then the states of $S \backslash S(f_*)$ are transient in the Markov chain $P(f_*)$ and policy f_*^∞ is an average optimal policy. However, we may not be so fortunate in our selection of f_*^∞. In that case, Derman suggests the following approach to find an optimal policy (see pp. 76–78 in Derman's book 1970). Let S_1 be defined by

$$S_1 = \left\{ i \;\middle|\; \exists a \in A(i) \text{ such that } \begin{cases} \sum_j \{\delta_{ij} - p_{ij}(a)\} v_j = 0 \\ v_i + \sum_j \{\delta_{ij} - p_{ij}(a)\} u_j = r_i(a) \end{cases} \right\}. \tag{5}$$

By Lemma 1, $S \backslash S_1$ must consist entirely of transient states under every optimal policy. Let S_2 be defined by

$$S_2 = \left\{ i \in S_1 \;\middle|\; \begin{array}{l} \exists a \in A(i) \text{ with } \begin{cases} \sum_j \{\delta_{ij} - p_{ij}(a)\} v_j = 0 \\ v_i + \sum_j \{\delta_{ij} - p_{ij}(a)\} u_j = r_i(a) \end{cases} \\ \text{which satisfies } p_{ij}(a) = 0 \text{ for all } j \in S \backslash S_1 \end{array} \right\}. \tag{6}$$

Also by Lemma 1, the states of $S_1 \backslash S_2$ must be transient under at least one optimal policy f_*^∞. Let S_3 and $A_3(i)$, $i \in S_3$ be defined as

$$S_3 = S \backslash S_2; \qquad A_3(i) = \left\{ a \in A(i) \;\middle|\; \sum_j \{\delta_{ij} - p_{ij}(a)\}\phi_j = 0 \right\}, \quad i \in S_3. \tag{7}$$

Consider the following linear program

$$\min \left\{ \sum_{j \in S_3} w_j \;\middle|\; \sum_{j \in S_3} \{\delta_{ij} - p_{ij}(a)\} w_j \geq s_i(a), i \in S_3,\ a \in A_3(i) \right\}, \tag{8}$$

where $s_i(a) = r_i(a) - \sum_{j \notin S_3} \{\delta_{ij} - p_{ij}(a)\} u_j^* - \phi_i$.

Theorem 5

(1) *The linear program* (8) *has a finite optimal solution.*
(2) *Let* w^* *be an optimal solution of* (8)*. Then, for each* $i \in S_3$ *there exists at least one action* $f_*(i)$ *satisfying* $\sum_{j\in S_3}\{\delta_{ij} - p_{ij}(f_*)\}w_j^* = s_i(f_*)$.
(3) *Let* f_*^∞ *be such that*

$$\begin{cases} \sum_j \{\delta_{ij} - p_{ij}(f_*)\}\phi_j = 0, & i \in S_2 \\ \phi_i + \sum_j \{\delta_{ij} - p_{ij}(f_*)\}u_j^* = r_i(f_*), & i \in S_2 \end{cases}$$

and $\sum_{j\in S_3}\{\delta_{ij} - p_{ij}(f_*)\}w_j^* = s_i(f_*)$, $i \in S_3$. *Then,* f_*^∞ *is an average optimal policy.*

Hence, in order to find an optimal policy in the multichain case, by the results of Denardo and Fox (1968) and Derman (1970), one has to execute the following procedure:

1. Determine an optimal solution (v^*, u^*) of the linear program (3) to find the value vector $\phi = v^*$.
2. Determine, by (5), (6) and (7), the sets S_1, S_2, S_3 and $A_3(i)$, $i \in S_3$.
3. Compute $s_i(a) = r_i(a) - \sum_{j\notin S_3}\{\delta_{ij} - p_{ij}(a)\}u_j^* - \phi_i$, $i \in S_3$, $a \in A_3(i)$.
4. Determine an optimal solution w^* of the linear program (8).
5. Determine an optimal policy f_*^∞ as described in Theorem 5.

This rather complicated approach elicited from Derman the remark (see Derman 1970, p. 84): "No satisfactory treatment of the dual program for the multiple class case has been published", which was for Hordijk and myself the reason to start research on this topic. In Hordijk and Kallenberg (1979) the following result was proved.

Theorem 6 *Let* (x^*, y^*) *be an extreme optimal solution of the dual program* (4)*. Then, any stationary deterministic policy* f_*^∞ *such that*

$$\begin{cases} x_i^*(f_*(i)) > 0 & \text{if } i \in S_* \\ y_i^*(f_*(i)) > 0 & \text{if } i \notin S_* \end{cases}, \quad \text{where } S_* = \left\{ i \mid \sum_a x_i^*(a) > 0 \right\},$$

is well-defined and is an average optimal policy.

This result is based on the following propositions, where:

– Proposition 1 is related to Lemma 1;
– Proposition 2 is related to the definitions of S_2;
– Proposition 3 is related to Lemma 2; it also uses the property that the columns of positive variables of an extreme optimal solution are linearly independent.

Proposition 1 *Let* $(v^* = \phi, u^*)$ *be an optimal solution of program* (3)*. Then,*

$$\begin{cases} \sum_j \{\delta_{ij} - p_{ij}(f_*)\}\phi_j = 0, & i \in S \\ \phi_i + \sum_j \{\delta_{ij} - p_{ij}(f_*)\}u_j^* = r_i(f_*), & i \in S_*. \end{cases}$$

Proposition 2 *The subset* S_* *of* S *is closed in the Markov chain* $P(f_*)$.

Proposition 3 *The states of* $S\backslash S_*$ *are transient in the Markov chain* $P(f_*)$.

The correspondence between feasible solutions (x, y) of (4) and randomized stationary policies π^∞ is given by the following mappings. For a feasible solution (x, y) the corresponding policy $\pi^\infty(x, y)$ is defined by

$$\pi_{ia}(x, y) = \begin{cases} \frac{x_i(a)}{\sum_a x_i(a)} & \text{if } \sum_a x_i(a) > 0 \\ \frac{y_i(a)}{\sum_a y_i(a)} & \text{if } \sum_a x_i(a) = 0. \end{cases} \tag{9}$$

Conversely, for a stationary policy π^∞, we define a feasible solution (x^π, y^π) of the dual program (4) by

$$\begin{cases} x_i^\pi(a) = \{\sum_j \beta_j \{P^*(\pi)\}_{ji}\} \cdot \pi_i(a) \\ y_i^\pi(a) = \{\sum_j \beta_j \{D(\pi)\}_{ji} + \sum_j \gamma_j \{P^*(\pi)\}_{ji}\} \cdot \pi_i(a), \end{cases} \tag{10}$$

where $P^*(\pi)$ and $D(\pi)$ are the stationary and the deviation matrix of the transition matrix $P(\pi)$; $\gamma_j = 0$ on the transient states and constant on each recurrent class under $P(\pi)$ (for the precise definition of γ see Hordijk and Kallenberg 1979).

Now, we will present some examples which show the essential difference between irreducible, unichained and multichained MDPs.

Example 2 It is well-known that in the irreducible case each extreme optimal solution has exactly one positive x-variable. It is also well known that in other cases some states can have no positive x-variables, i.e., S_* is a proper subset of S.

This example shows an MDP with an extreme optimal solution which has two positive x-variables for some state. Hence, the two corresponding deterministic policies, which can constructed via Theorem 6, are both optimal.

Furthermore, this extreme feasible solution is mapped on a nondeterministic policy. Let $S = \{1, 2, 3\}$; $A(1) = \{1\}$, $A(2) = \{1\}$, $A(3) = \{1, 2\}$; $r_1(1) = 1$, $r_2(1) = 2$, $r_3(1) = 4$, $r_3(2) = 3$; $p_{13}(1) = p_{23}(1) = p_{31}(1) = p_{32}(2) = 1$ (other transitions are 0).

The dual program (4) of this MDP is (take $\beta_1 = \beta_2 = \frac{1}{4}$, $\beta_3 = \frac{1}{2}$):

$$\text{maximize } x_1(1) + 2x_2(1) + 4x_3(1) + 3x_3(2)$$

subject to

$$\begin{aligned}
x_1(1) - x_3(1) &= 0 \\
x_2(1) - x_3(2) &= 0 \\
-x_1(1) - x_2(1) + x_3(1) + x_3(2) &= 0 \\
x_1(1) + y_1(1) - y_3(1) &= \tfrac{1}{4} \\
x_2(1) + y_2(1) - y_3(2) &= \tfrac{1}{4} \\
x_3(1) + x_3(2) - y_1(1) - y_2(1) + y_3(1) + y_3(2) &= \tfrac{1}{2} \\
x_1(1), x_2(1), x_3(1), x_3(2), y_1(1), y_2(1), y_3(1), y_3(2) &\geq 0
\end{aligned}$$

The feasible solution (x, y), where $x_1(1) = x_2(1) = x_3(1) = x_3(2) = \frac{1}{4}$, $y_1(1) = y_2(1) = y_3(1) = y_3(2) = 0$, is an extreme optimal solution. Observe that state 3 has two positive x-variables.

Example 3 This example shows that the mapping (9) is not a bijective mapping. Let $S = \{1, 2, 3, 4\}$; $A(1) = \{1\}$, $A(2) = \{1, 2\}$, $A(3) = \{1, 2\}$, $A(4) = \{1\}$; $p_{12}(1) = p_{23}(1) =$

$p_{24}(2) = p_{33}(1) = p_{31}(2) = p_{44}(1) = 1$ (other transitions are 0). Since the rewards are not important for this property, we have omitted these numbers.

The constraints of the dual program are (take $\beta_j = \frac{1}{4}$, $1 \le j \le 4$):

$$\begin{aligned}
x_1(1) - x_3(2) &= 0\\
-x_1(1) + x_2(1) + x_2(2) &= 0\\
-x_2(1) + x_3(2) &= 0\\
-x_2(2) &= 0\\
x_1(1) + y_1(1) - y_3(2) &= \tfrac{1}{4}\\
x_2(1) + x_2(2) - y_1(1) + y_2(1) + y_2(2) &= \tfrac{1}{4}\\
x_3(1) + x_3(2) - y_2(1) + y_3(2) &= \tfrac{1}{4}\\
x_4(1) - y_2(2) &= \tfrac{1}{4}\\
x_1(1),\ x_2(1),\ x_2(2),\ x_3(1),\ x_3(2),\ x_4(1),\ y_1(1),\ y_2(1),\ y_2(2),\ y_3(2) &\ge 0
\end{aligned}$$

First, consider the feasible solution (x^1, y^1) with $x_1^1(1) = x_2^1(1) = \frac{1}{4}$, $x_2^1(2) = x_3^1(1) = 0$, $x_3^1(2) = x_4^1(1) = \frac{1}{4}$; $y_1^1(1) = y_2^1(1) = y_2^1(2) = y_3^1(2) = 0$. This feasible solution is mapped on the deterministic policy f_1^∞ with $f_1(1) = f_1(2) = 1$, $f_1(3) = 2$, $f_1(4) = 1$.

Then, consider the feasible solution (x^2, y^2) with $x_1^2(1) = x_2^2(1) = \frac{1}{6}$, $x_2^2(1) = x_3^2(1) = 0$, $x_3^2(2) = \frac{1}{6}$, $x_4^2(1) = \frac{1}{2}$, $y_1^2(1) = \frac{1}{6}$, $y_2^2(1) = 0$, $y_2^2(2) = \frac{1}{4}$, $y_3^2(2) = \frac{1}{12}$. This feasible solution is mapped on the deterministic policy f_2^∞ with $f_2(1) = f_2(2) = 1$, $f_2(3) = 2$, $f_2(4) = 1$. Notice that $(x^1, y^1) \neq (x^2, y^2)$ and $f_1^\infty = f_2^\infty$.

Example 4 This example shows that a feasible nonoptimal solution can be mapped on an optimal policy. Let $S = \{1, 2, 3\}$; $A(1) = A(2) = \{1, 2\}$, $A(3) = \{1\}$; $p_{12}(1) = p_{13}(2) = p_{21}(1) = p_{22}(2) = p_{33}(1) = 1$ (other transitions are 0); $r_1(1) = 1$, $r_1(2) = r_2(1) = r_2(2) = r_3(1) = 0$.

The dual program for this model is (take $\beta_1 = \beta_2 = \beta_3 = \frac{1}{3}$):

maximize $x_1(1)$

subject to

$$\begin{aligned}
x_1(1) + x_1(2) - x_2(1) &= 0\\
x_2(1) - x_1(1) &= 0\\
-x_1(2) &= 0\\
x_1(1) + x_1(2) + y_1(1) + y_1(2) - y_2(1) &= \tfrac{1}{3}\\
x_2(1) + x_2(2) - y_1(1) + y_2(1) &= \tfrac{1}{3}\\
x_3(1) - y_1(2) &= \tfrac{1}{3}\\
x_1(1),\ x_1(2),\ x_2(1),\ x_2(2),\ x_3(1),\ y_1(1),\ y_1(2),\ y_2(1) &\ge 0
\end{aligned}$$

The solution (x, y) given by $x_1(1) = \frac{1}{6}$, $x_1(2) = 0$, $x_2(1) = \frac{1}{6}$, $x_2(2) = 0$, $x_3(1) = \frac{2}{3}$, $y_1(1) = 0$, $y_1(2) = \frac{1}{3}$, $y_2(1) = \frac{1}{6}$ is a feasible solution, but not an optimal solution. Notice that $x_1^*(1) = x_2^*(1) = x_3^*(1) = \frac{1}{3}$ and all other variables 0 is an optimal solution and that the x-part of the optimal solution is unique. However, the policy f^∞ which corresponds to (x, y) has $f(1) = f(2) = f(3) = 1$ and is an optimal policy.

Example 5 In this last example, we show that the general unichain case needs an approach different from the unichain case; even the additional search procedure is not sufficient. In the general unichain case the value vector is a constant vector and the linear programs (1) and (2) may be considered. Let $S = \{1, 2, 3\}$; $A(1) = \{1\}$, $A(2) = A(3) = \{1, 2\}$; $r_1(1) = r_2(1) = 0$, $r_2(2) = r_3(1) = 1$, $r_3(2) = 0$; $p_{12}(1) = p_{21}(1) = p_{22}(2) = p_{33}(1) = p_{32}(2) = 1$ (other transitions are 0). This is a general unichained MDP, because the policy f^∞ with $f(1) = 1$, $f(2) = f_*(3) = 2$ is an optimal policy and has a single chain structure. The dual program (2) of this model is:

$$\text{maximize } x_2(2) + x_3(1)$$

subject to

$$\begin{aligned}
x_1(1) - x_2(1) &= 0\\
-x_1(1) + x_2(1) - x_3(2) &= 0\\
-x_3(2) &= 0\\
x_1(1) + x_2(1) + x_2(2) + x_3(1) + x_3(2) &= 1\\
x_1(1),\ x_2(1),\ x_2(2),\ x_3(1),\ x_3(1) &\geq 0
\end{aligned}$$

x given by $x_1(1) = x_2(1) = x_2(2) = x_3(2) = 0$, $x_3(1) = 1$ is an extreme optimal solution. In state 3, the policy corresponding to x chooses action 1. The choice in state 2 for an optimal policy has to be action 2. Since the set of the states 1 and 2 is closed under any policy, it is impossible to search for actions in these states with transitions to state 3.

4 State-action frequencies and problems with constraints

4.1 Introduction

"State-action frequencies and problems with constraints" is the title of chapter 7 of Derman's book. This chapter may be concerned as the starting point for the study of MDPs with additional constraints. In such problems it is not obvious that optimal policies exist. It is also not necessarily true that optimal policies, if they exist, belong to the class $C(D)$ or $C(S)$.

MDPs with additional constraints occur in a natural way in all kind of applications. For instance in inventory management, where one wants to minimize the total costs under the constraint that the shortage is bounded by a given number.

In general, for MDPs with additional constraints, a policy which is optimal simultaneously for all starting states does not exist. Therefore, we consider problems with a given *initial distribution* β, i.e., β_j is a given probability that state j is the starting state. A special case is $\beta_j = 1$ for $j = i$ and $\beta_j = 0$ for $j \neq i$, i.e., that state i is the (fixed) starting state.

In many cases reward and cost functions are specified in terms of expectations of some function of the *state-action frequencies*. Given the initial distribution β, we define for any policy R, any time point t and any state-action pair $(i, a) \in S \times A$, the action-state frequency $x_{ia}^R(t)$ by

$$x_{ia}^R(t) = \sum_{j \in S} \beta_j \cdot \mathbb{P}_R\{X_t = i, Y_t = a \mid X_1 = j\}. \tag{11}$$

For the additional constraints we assume that, besides the immediate rewards $r_i(a)$, there are also certain immediate costs $c_i^k(a)$, $i \in S$, $a \in A(i)$ for $k = 1, 2, \ldots, m$.

Let β be an arbitrary initial distribution. For any policy R, let the average reward and the k-th average cost function with respect to the initial distribution β be defined by

$$\phi(\beta, R) = \liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{j\in S} \beta_j \cdot \sum_{i,a} \mathbb{P}_R\{X_t = i, Y_t = a \mid X_1 = j\} \cdot r_i(a) \tag{12}$$

and

$$c^k(\beta, R) = \liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{j\in S} \beta_j \cdot \sum_{i,a} \mathbb{P}_R\{X_t = i, Y_t = a \mid X_1 = j\} \cdot c_i^k(a). \tag{13}$$

A policy R is a feasible policy for a *constrained Markov decision problem*, shortly CMDP, if the k-the cost function is bounded by a given number b_k for $k = 1, 2, \ldots, m$, i.e., if $c^k(\beta, R) \le b_k$, $k = 1, 2, \ldots, m$.

An *optimal policy* R^* for this criterion is a feasible policy that maximizes $\phi(\beta, R)$, i.e.,

$$\phi(\beta, R^*) = \sup_R\{\phi(\beta, R) \mid c^k(\beta, R) \le b_k, k = 1, 2, \ldots, m\}. \tag{14}$$

For any policy R and any $T \in \mathbb{N}$, we denote the *average expected state-action frequencies in the first T periods* by

$$x_{ia}^T(R) = \frac{1}{T} \sum_{t=1}^{T} x_{ia}^R(t), \quad (i, a) \in S \times A. \tag{15}$$

By $X(R)$ we denote the limit points of the vectors $\{x^T(R), T = 1, 2, \ldots\}$. For any $T \in \mathbb{N}$, $x^T(R)$ satisfies $\sum_{(i,a)} x_{ia}^T(R) = 1$; so also $\sum_{(i,a)} x_{ia}(R) = 1$ for all $x(R) \in X(R)$.

Since $\mathbb{P}_{\pi^\infty}\{X_t = i, Y_t = a \mid X_1 = j\} = \{P^{t-1}(\pi)\}_{ji} \cdot \pi_{ia}$, $(i, a) \in S \times A$ for all $\pi^\infty \in C(S)$, we have $\lim_{T\to\infty} x_{ia}^T(\pi^\infty) = \sum_{j\in S} \beta_j\{P^*(\pi)\}_{ji} \cdot \pi_{ia}$, i.e., $X(\pi^\infty)$ consists of only one element, namely the vector $x(\pi)$, where $x_{ia}(\pi) = \{\beta^T P^*(\pi)\}_i \cdot \pi_{ia}$, $(i, a) \in S \times A$.

Let the policy set C_1 be the set of *convergent policies*, defined by

$$C_1 = \{R \mid X(R) \text{ consists of one element}\}. \tag{16}$$

Hence, $C(S) \subseteq C_1$. Furthermore, define the vector sets L, $L(M)$, $L(C)$, $L(S)$ and $L(D)$ by

$$\begin{aligned}
L &= \{x(R) \in X(R) \mid R \text{ is an arbitrary policy}\};\\
L(M) &= \{x(R) \in X(R) \mid R \text{ is a Markov policy}\};\\
L(C) &= \{x(R) \in X(R) \mid R \text{ is a convergent policy}\};\\
L(S) &= \{x(R) \in X(R) \mid R \text{ is a stationary policy}\};\\
L(D) &= \{x(R) \in X(R) \mid R \text{ is a deterministic policy}\}.
\end{aligned}$$

The following result is due to Derman (1970, pp. 93–94).

Theorem 7 *$L = L(M) = \overline{L(S)} = \overline{L(D)}$, where $\overline{L(S)}$ and $\overline{L(D)}$ are the closed convex hull of the sets $L(S)$ and $L(D)$, respectively.*

4.2 The unichain case

Derman has also shown (Derman 1970, pp. 95–96) that in the unichain case a feasible CMDP has an optimal *stationary* policy. He showed that $L(S) = X$, where

$$X = \left\{ x \in \mathbb{R}^{|S\times A|} \;\middle|\; \begin{array}{ll} \sum_{i,a}\{\delta_{ij} - p_{ij}(a)\}x_i(a) = 0, & j \in S \\ \sum_{i,a} x_i(a) = 1 & \\ x_i(a) \geq 0, & i \in S,\ a \in A(i) \end{array} \right\}. \tag{17}$$

Since X is a closed convex set, this result also implies that $L(S) = \overline{L(S)}$. Hence, the CMDP (14) can be solved by the following algorithm.

Algorithm 1

1. Determine an optimal solution x^* of the linear program

$$\max \left\{ \sum_{i,a} r_i(a)x_i(a) \;\middle|\; \begin{array}{ll} \sum_{i,a}\{\delta_{ij} - p_{ij}(a)\}x_i(a) = 0, & j \in S \\ \sum_{i,a} x_i(a) = 1 & \\ \sum_{i,a} c_i^k(a)x_i(a) \leq b_k, & k = 1, 2, \ldots, m \\ x_i(a) \geq 0, & (i,a) \in S \times A \end{array} \right\}. \tag{18}$$

   (if (18) is infeasible, then problem (14) is also infeasible).
2. Take

$$\pi_{ia}^* = \begin{cases} x_i^*(a)/x_i^*, & a \in A(i), i \in S_* \\ \text{arbitrary} & \text{otherwise,} \end{cases}$$

   where $x_i^* = \sum_a x_i^*(a)$ and $S_* = \{i \mid x_i^* > 0\}$.

4.3 The multichain case

The multichain case was solved by Hordijk and Kallenberg (see Kallenberg 1980, 1983 and Hordijk and Kallenberg 1984). First, they generalized Theorem 7 in the following way.

Theorem 8 $L = L(M) = L(C) = \overline{L(S)} = \overline{L(D)}$.

Then, they showed that $L = XY$, where

$$XY = \left\{ x \;\middle|\; \exists y \text{ s.t. } \begin{array}{ll} \sum_{i,a}\{\delta_{ij} - p_{ij}(a)\}x_{ia} = 0, & j \in S \\ \sum_a x_{ja} + \sum_{i,a}\{\delta_{ij} - p_{ij}(a)\}y_{ia} = \beta_j, & j \in S \\ x_{ia},\ y_{ia} \geq 0, & (i,a) \in S \times A \end{array} \right\}. \tag{19}$$

From the above results it follows that any extreme point of XY is an element of $L(D)$. The next example shows the converse statement is not true, in general.

Example 6 Take the MDP with $S = \{1,2,3\}$; $A(1) = \{1,2\}$, $A(2) = \{1,2\}$, $A(3) = \{1\}$; $p_{12}(1) = p_{13}(2) = p_{22}(1) = p_{21}(2) = p_{33}(1) = 1$ (other transitions are 0). Since the rewards are not important for this property, we have omitted these numbers. Let $\beta_1 = \beta_2 = \beta_3 = \frac{1}{3}$. Consider f_1^∞, f_2^∞, f_3^∞, where $f_1(1) = 2$, $f_1(2) = 1$, $f_1(3) = 1$; $f_2(1) = 2$, $f_2(2) = 2$, $f_2(3) = 1$; $f_3(1) = 1$, $f_3(2) = 1$, $f_3(3) = 1$.

For these policies one easily verifies that:

$$x_{11}(f_1^\infty)=0,\quad x_{12}(f_1^\infty)=0,\quad x_{21}(f_1^\infty)=\frac{1}{3},\quad x_{22}(f_1^\infty)=0,\quad x_{31}(f_1^\infty)=\frac{2}{3};$$
$$x_{11}(f_2^\infty)=0,\quad x_{12}(f_2^\infty)=0,\quad x_{21}(f_2^\infty)=0,\quad x_{22}(f_2^\infty)=0,\quad x_{31}(f_2^\infty)=1;$$
$$x_{11}(f_3^\infty)=0,\quad x_{12}(f_3^\infty)=0,\quad x_{21}(f_3^\infty)=\frac{2}{3},\quad x_{22}(f_1^\infty)=0,\quad x_{31}(f_1^\infty)=\frac{1}{3}.$$

Since $x(f_1^\infty)=\frac{1}{2}x(f_2^\infty)+\frac{1}{2}x(f_3^\infty)$, $x(f_1^\infty)$ is not an extreme point of XY.

In order to solve the CMDP (14) we consider the linear program

$$\max\left\{\sum_{i,a} r_i(a)x_i(a)\;\middle|\;\begin{array}{ll}\sum_{i,a}\{\delta_{ij}-p_{ij}(a)\}x_i(a)=0, & j\in S\\ \sum_a x_j(a)+\sum_{i,a}\{\delta_{ij}-p_{ij}(a)\}y_i(a)=\beta_j, & j\in S\\ \sum_{i,a} c_i^k(a)x_i(a)\le b_k, & 1\le k\le m\\ x_i(a),\,y_i(a)\ge 0, & (i,a)\in S\times A\end{array}\right\}. \tag{20}$$

The next theorem shows how an optimal policy for the CMDP (14) can be computed. This policy may lie outside the set of stationary policies.

Theorem 9

(1) *Problem* (14) *is feasible if and only if problem* (20) *is feasible.*
(2) *The optima of* (14) *and* (20) *are equal.*
(3) *If R is optimal for problem* (14)*, then $x(R)$ is optimal for* (20).
(4) *Let (x,y) be an optimal solution of problem* (20) *and let $x=\sum_{k=1}^n p_k x(f_k)$, where $p_k\ge 0$ and $\sum_{k=1}^n p_k=1$ and $C(D)=\{f_1^\infty,f_2^\infty,\ldots,f_n^\infty\}$. Let $R\in C(M)$ such that $\sum_j \beta_j\cdot\mathbb{P}_R\{X_t=i,Y_t=a\mid X_1\}=\sum_j\beta_j\cdot\sum_k p_k\cdot\mathbb{P}_{f_k^\infty}\{X_t=i,Y_t=a\mid X_1\}=\beta_j\}$ for all $(i,a)\in S\times A$ and all $t\in\mathbb{N}$. Then, R is an optimal solution of problem* (14).

To compute an optimal policy from an optimal solution (x,y) of the linear program (20), we first have to express x as $x=\sum_{k=1}^n p_k x(f_k^\infty)$, where $p_k\ge 0$ and $\sum_{k=1}^n p_k=1$. Next, we have to determine the policy $R=(\pi^1,\pi^2,\ldots)\in C(M)$ such that R satisfies $\sum_j\beta_j\times\mathbb{P}_R\{X_t=i,Y_t=a|X_1\}=\sum_j\beta_j\cdot\sum_k p_k\cdot\mathbb{P}_{f_k^\infty}\{X_t=i,Y_t=a|X_1\}=\beta_j\}$ for all $(i,a)\in S\times A$ and all $t\in\mathbb{N}$. The decision rules π^t, $t\in\mathbb{N}$, can be determined by

$$\pi_{ia}^t=\begin{cases}\frac{\sum_j\beta_j\cdot\sum_k p_k\{P^{t-1}(f_k)\}_{ji}\cdot\delta_{af_k(i)}}{\sum_j\beta_j\cdot\sum_k p_k\{P^{t-1}(f_k)\}_{ji}} & \text{if } \sum_j\beta_j\cdot\sum_k p_k\{P^{t-1}(f_k)\}_{ji}\neq 0\\ \text{arbitrary} & \text{if } \sum_j\beta_j\cdot\sum_k p_k\{P^{t-1}(f_k)\}_{ji}=0.\end{cases}$$

Hence, the following algorithm constructs a policy $R\in C(M)\cap C_1$ which is optimal for CMDP problem (14).

Algorithm 2

1. Determine an optimal solution (x^*,y^*) of linear program (20) (if (20) is infeasible, then problem (14) is also infeasible).
2. (a) Let $C(D)=\{f_1^\infty,f_2^\infty,\ldots,f_n^\infty\}$ and compute $P^*(f_k)$ for $k=1,2,\ldots,n$.
   (b) Take

$$x_{ia}^k=\begin{cases}\sum_j\beta_j\cdot\{P^*(f_k)\}_{ji} & a=f_k(i)\\ 0 & a\neq f_k(i)\end{cases},\qquad i\in S,\ k=1,2,\ldots,n.$$

3. Determine p_k, $k = 1, 2, \ldots, n$ as feasible solution of the linear system

$$\begin{cases} \sum_{k=1}^{n} p_k x_{ia}^k = x_{ia}^*, & a \in A(i), i \in S \\ \sum_{k=1}^{n} p_k = 1 & \\ p_k \geq 0 & k = 1, 2, \ldots, n \end{cases}$$

4. $R = (\pi^1, \pi^2, \ldots)$, defined by

$$\pi_{ia}^t = \begin{cases} \frac{\sum_j \beta_j \cdot \sum_k p_k \{P^{t-1}(f_k)\}_{ji} \cdot \delta_{af_k(i)}}{\sum_j \beta_j \cdot \sum_k p_k \{P^{t-1}(f_k)\}_{ji}} & \text{if } \sum_j \beta_j \cdot \sum_k p_k \{P^{t-1}(f_k)\}_{ji} \neq 0 \\ \text{arbitrary} & \text{if } \sum_j \beta_j \cdot \sum_k p_k \{P^{t-1}(f_k)\}_{ji} = 0 \end{cases}$$

is an optimal policy for problem (14).

In the next example Algorithm 2 is applied on a CMDP.

Example 7 Let $S = \{1, 2, 3\}$; $A(1) = \{1, 2\}$, $A(2) = \{1\}$, $A(3) = \{1, 2\}$; $p_{12}(1) = p_{13}(2) = p_{22}(1) = p_{33}(1) = p_{32}(2) = 1$ (other transitions are 0); $r_1(1) = 0$, $r_1(2) = 0$, $r_2(1) = 1$, $r_3(1) = r_3(2) = 0$; $\beta_1 = \frac{1}{4}$, $\beta_2 = \frac{3}{16}$, $\beta_3 = \frac{9}{16}$. As constraints we have bounds for the value $x_{21}(R) : \frac{1}{4} \leq x_{21}(R) \leq \frac{1}{2}$. If we apply Algorithm 2 we obtain the following.

maximize $x_2(1)$

subject to

$$\begin{aligned}
x_1(1) + x_1(2) & & & & & & & = & 0 \\
-x_1(1) & & -x_3(2) & & & & & = & 0 \\
-x_1(2) & & +x_3(2) & & & & & = & 0 \\
x_1(1) + x_1(2) & & & & +y_1(1) + y_1(2) & & & = & \tfrac{1}{4} \\
& x_2(1) & & & -y_1(1) & & -y_3(2) & = & \tfrac{3}{16} \\
& & x_3(1) + x_3(2) & & & & -y_1(2) + y_3(2) & = & \tfrac{9}{16} \\
& x_2(1) & & & & & & \leq & \tfrac{1}{2} \\
& -x_2(1) & & & & & & \leq & -\tfrac{1}{4}
\end{aligned}$$

$$x_1(1),\ x_1(2),\ x_2(1),\ x_3(1),\ x_3(2),\ y_1(1),\ y_1(2),\ y_3(2) \geq 0$$

with optimal solution: $x_1^*(1) = 0$, $x_1^*(2) = 0$, $x_2^*(1) = \frac{1}{2}$, $x_3^*(1) = \frac{1}{2}$, $x_3^*(2) = 0$; $y_1^*(1) = 0$, $y_1^*(2) = \frac{1}{4}$, $y_3^*(2) = \frac{5}{16}$.

There are four deterministic policies:

$$\begin{aligned}
&f_1(1) = 1, \quad f_1(2) = 1, \quad f_1(3) = 1; \quad f_2(1) = 1, \quad f_2(2) = 1, \quad f_2(3) = 2; \\
&f_3(1) = 2, \quad f_3(2) = 1, \quad f_3(3) = 1; \quad f_4(1) = 2, \quad f_4(2) = 1, \quad f_4(3) = 2.
\end{aligned}$$

The corresponding vectors x^1, x^2, x^3, x^4 are:

$$\begin{aligned}
&x_1^1(1) = 0; \quad x_1^1(2) = 0; \quad x_2^1(1) = \frac{7}{16}; \quad x_3^1(1) = \frac{9}{16}; \quad x_3^1(2) = 0. \\
&x_1^2(1) = 0; \quad x_1^2(2) = 0; \quad x_2^2(1) = 1; \quad x_3^2(1) = 0; \quad x_3^2(2) = 0.
\end{aligned}$$

$$x_1^3(1)=0;\quad x_1^3(2)=0;\quad x_2^3(1)=\frac{3}{16};\quad x_3^3(1)=\frac{13}{16};\quad x_3^3(2)=0.$$
$$x_1^4(1)=0;\quad x_1^4(2)=0;\quad x_2^4(1)=1;\quad x_3^4(1)=0;\quad x_3^4(2)=0.$$

For the numbers $p_1, p_2, p_3, p_4 \geq 0$ such that $p_1x^1 + p_2x^2 + p_3x^3 + p_4x^4 = x^*$ and $\sum_{k=1}^4 p_k = 1$, we obtain: $p_1 = \frac{8}{9}$, $p_2 = \frac{1}{9}$, $p_3 = 0$, $p_4 = 0$.

Since

$$P^t(f_1)=\begin{pmatrix}0&1&0\\0&1&0\\0&0&1\end{pmatrix}\quad\text{and}\quad P^t(f_2)=\begin{pmatrix}0&1&0\\0&1&0\\0&1&0\end{pmatrix}\quad\text{for all } t\in\mathbb{N},$$

we obtain $R=(\pi^1,\pi^2,\dots)$ with $\pi_{11}^t=1,\ t\in\mathbb{N};\ \pi_{21}^t=1,\ t\in\mathbb{N};\ \pi_{31}^t=\left\{\begin{smallmatrix}\frac{8}{9} & t=1\\ 1 & t\geq 2\end{smallmatrix}\right.;\ \pi_{32}^t=\left\{\begin{smallmatrix}\frac{1}{9} & t=1\\ 0 & t\geq 2\end{smallmatrix}\right.$.

Remark Algorithm 2 is unattractive for practical problems. The number of calculations is prohibitive. Moreover, the use of Markov policies is inefficient in practice. Therefore, we also analyze the problem of finding an optimal stationary policy, if one exists.

For any feasible solution (x, y) of (20) we define a stationary policy $\pi^\infty(x, y)$ in a slightly different way as by (9). The difference is caused by the fact that for constrained MDPs β_j can be equal to zero in one or more states j, while in unconstrained MDPs we take $\beta_j > 0$ for all states j.

$$\pi_{ia}(x,y)=\begin{cases}x_i(a)/x_i & \text{if } \sum_a x_i(a)>0\\ y_i(a)/y_i & \text{if } \sum_a x_i(a)=0 \text{ and } \sum_a y_i(a)>0\\ \text{arbitrary} & \text{if } \sum_a x_i(a)=0 \text{ and } \sum_a y_i(a)=0.\end{cases}\tag{21}$$

In Kallenberg (1983) the following lemmata can be found.

Lemma 3 *If (x^*, y^*) is an optimal solution of problem (20) and the Markov chain $P(\pi(x^*, y^*))$ has one ergodic set plus a (perhaps empty) set of transient states, then $\pi^\infty(x^*, y^*)$ is an optimal policy for problem (14).*

Lemma 4 *If (x^*, y^*) is an optimal solution of problem (20) and x^* satisfies $x_i^*(a) = \pi_{ia}(x^*, y^*) \cdot \{\beta^T P^*(\pi(x^*, y^*))\}_i$ for all $(i, a) \in S \times A$, then $\pi^\infty(x^*, y^*)$ is an optimal policy for problem (14).*

Lemma 5 *If (x^*, y^*) is an optimal solution of problem (20) and furthermore $x_i^*(a)/x_i^* = y_i^*(a)/y_i^*$ for all pairs (i, a) with $i \in S_+$, $a \in A(i)$, where $x_i^* = \sum_a x_{ia}^*$, $y_i^* = \sum_a y_{ia}^*$ and $S_+ = \{i \mid x_i^* > 0, y_i^* > 0\}$, then the stationary policy $\pi^\infty(x^*, y^*)$ is an optimal policy for problem (14).*

The next example shows that for an optimal solution (x^*, y^*) of (20), the policy $\pi^\infty(x^*, y^*)$ is not an optimal solution of (14), even in the case that (14) has a stationary optimal policy.

Example 7 (continued)
Consider the MDP model of Example 7, but with as constraint $x_{21}(R) \leq \frac{1}{4}$. The linear program (20) for this constrained problem is:

$$
\begin{aligned}
&\text{maximize } x_2(1)\\
&\text{subject to}\\
&\begin{array}{rrrrrrrrrrrr}
x_1(1) &+ x_1(2) & & & & & & & & & = & 0\\
-x_1(1) & & & - x_3(2) & & & & & & & = & 0\\
& - x_1(2) & & + x_3(2) & & & & & & & = & 0\\
x_1(1) &+ x_1(2) & & & & + y_1(1) &+ y_1(2) & & & & = & \frac{1}{4}\\
& & x_2(1) & & & - y_1(1) & & - y_3(2) & & & = & \frac{3}{16}\\
& & & x_3(1) &+ x_3(2) & & - y_1(2) &+ y_3(2) & & & = & \frac{9}{16}\\
& & x_2(1) & & & & & & & & \leq & \frac{1}{4}
\end{array}\\
&x_1(1),\ x_1(2),\ x_2(1),\ x_3(1),\ x_3(2),\ y_1(1),\ y_1(2),\ y_3(2) \geq 0
\end{aligned}
$$

with optimal solution $x_1^*(1) = 0$, $x_1^*(2) = 0$, $x_2^*(1) = \frac{1}{4}$, $x_3^*(1) = \frac{3}{4}$, $x_3^*(2) = 0$; $y_1^*(1) = 0$, $y_1^*(2) = \frac{1}{4}$, $y_3^*(2) = \frac{1}{16}$ and with optimum value $\frac{1}{4}$. The corresponding stationary policy $\pi^\infty(x^*, y^*)$ gives $\pi_{12} = \pi_{21} = \pi_{31} = 1$, so this policy is in fact deterministic. This policy is not optimal, because $\phi(\pi^\infty(x^*, y^*)) = \frac{3}{16} < \frac{1}{4}$, the optimum of the linear program. Consider the stationary policy π^∞ with $\pi_{11} = \frac{1}{4}$, $\pi_{12} = \frac{3}{4}$, $\pi_{21} = \pi_{31} = 1$. For this policy we obtain $x_{12}(\pi^\infty) = \frac{1}{4}$ and $\phi(\pi^\infty) = \frac{1}{4}$, the optimum value of the linear program. So, this policy is feasible and optimal.

If the conditions of Lemma 5 are not satisfied, we can try to find for the same x^* another y^*, say $\overline{y}$, such that $(x^*, \overline{y})$ is feasible for (20), and consequently also optimal, and satisfies the conditions of Lemma 5. To achieve this, we need $\overline{y}_i(a)/\overline{y}_i = \pi_{ia}$, $a \in A(i)$, $i \in \{j \mid x_j^* > 0, \overline{y}_j > 0\}$, which is equivalent to $\overline{y}_i(a) = \overline{y}_i \cdot \pi_{ia}$, $a \in A(i)$, $i \in \{j \mid x_j^* > 0\}$. Hence, $\overline{y}$ has to satisfy the following linear system in the y-variables (x^* is fixed)

$$
\begin{cases}
\sum_{i \notin S_*} \sum_a \{\delta_{ij} - p_{ij}(a)\}\overline{y}_i(a) + \sum_{i \in S_*} \{\delta_{ij} - p_{ij}(\pi)\}\overline{y}_i = \beta_j - x_j^*, & j \in S\\
\overline{y}_i(a) \geq 0, i \notin S_*, a \in A(i); \overline{y}_i \geq 0, i \in S_*, & \text{with } S_* = \{j \mid \sum_a x_j^*(a) > 0\}.
\end{cases}
\tag{22}
$$

Example 7 (continued)
The optimal solution (x^*, y^*) with $x_1^*(1) = 0$, $x_1^*(2) = 0$, $x_2^*(1) = \frac{1}{4}$, $x_3^*(1) = \frac{3}{4}$, $x_3^*(2) = 0$; $y_1^*(1) = 0$, $y_1^*(2) = \frac{1}{4}$, $y_3^*(2) = \frac{1}{16}$ does not satisfy $x_i^*(a)/x_i^* = y_i^*(a)/y_i^*$ for all $a \in A(i)$, $i \in S_+$, because $S_+ = \{3\}$ and $x_3^*(2)/x_3^* = 0$ and $y_3^*(2)/y_3^* = 1$. The system (22) becomes $\overline{y}_1(1) + \overline{y}_1(2) = \frac{4}{16}$; $-\overline{y}_1(1) = -\frac{1}{16}$; $-\overline{y}_1(2) = -\frac{3}{16}$; $\overline{y}_1(1), \overline{y}_1(2) \geq 0$. This system has the solution $\overline{y}_1(1) = \frac{1}{16}$, $\overline{y}_1(2) = \frac{3}{16}$. The stationary policy π^∞ with $\pi_{11} = \frac{1}{4}$, $\pi_{12} = \frac{3}{4}$, $\pi_{21} = \pi_{31} = 1$ is optimal for problem (14).

Remark If the x-part of problem (20) is unique and (22) is infeasible, then problem (14) has no optimal stationary policy. If the x-part of problem (20) is not unique and (22) is infeasible, then it is still possible that there exists an optimal stationary policy. In that case we can compute every extreme optimal solution of the linear program (20), and for each of these extreme optimal solutions we can perform the above analysis in order to search for an optimal stationary policy. We show an example of this approach.

Example 8 Take the MDP with $S = \{1, 2, 3\}$; $A(1) = \{1, 2\}$, $A(2) = \{1, 2\}$, $A(3) = \{1\}$; $p_{12}(1) = p_{13}(2) = p_{22}(1) = p_{21}(2) = p_{33}(1) = 1$ (other transitions are 0). $r_1(1) = r_1(2) = 0$, $r_2(1) = 1$, $r_2(2) = 0$, $r_3(1) = 1$. Let $\beta_1 = \beta_2 = \beta_3 = \frac{1}{3}$. Add as only constraint $x_{21}(R) \geq \frac{1}{9}$. The formulation of the linear program (20) becomes:

$$\text{maximize } x_2(1) + x_3(1)$$

subject to

$$\begin{array}{rrrrrrrrr}
x_1(1) + x_1(2) & - x_2(2) & & & & & = & 0 \\
- x_1(1) & + x_2(2) & & & & & = & 0 \\
- x_1(2) & & & & & & = & 0 \\
x_1(1) + x_1(2) & & & + y_1(1) + y_1(2) & - y_2(2) & = & \frac{1}{3} \\
& x_2(1) + x_2(2) & & - y_1(1) & + y_2(2) & = & \frac{1}{3} \\
& & x_3(1) & - y_1(2) & & = & \frac{1}{3} \\
& - x_2(1) & & & & \leq & -\frac{1}{9}
\end{array}$$

$$x_1(1),\ x_1(2),\ x_2(1),\ x_3(1),\ x_3(2),\ y_1(1),\ y_1(2),\ y_3(2) \geq 0$$

with extreme optimal solution $x_1^*(1) = 0$, $x_1^*(2) = 0$, $x_2^*(1) = \frac{1}{9}$, $x_2^*(2) = 0$, $x_3^*(1) = \frac{8}{9}$; $y_1^*(1) = 0$, $y_1^*(2) = \frac{5}{9}$, $y_2^*(2) = \frac{2}{9}$ and with optimum value 1. The x-part of this problem is not unique. It can easily be verified that $\hat{x}_1(1) = 0$, $\hat{x}_1(2) = 0$, $\hat{x}_2(1) = \frac{2}{3}$, $\hat{x}_2(2) = 0$, $\hat{x}_3(1) = \frac{1}{3}$; $\hat{y}_1(1) = \frac{1}{3}$, $\hat{y}_1(2) = 0$, $\hat{y}_2(2) = 0$ is also an extreme optimal solution. For the first extreme optimal solution (x^*, y^*) system (22) becomes

$$\overline{y}_1(1) + \overline{y}_1(2) = \frac{1}{3}; \quad -\overline{y}_1(1) = \frac{2}{9}; \quad \overline{y}_1(2) = -\frac{5}{9}; \quad \overline{y}_1(1), \overline{y}_1(2) \geq 0.$$

This system is obviously infeasible.

For the second extreme optimal solution $(\hat{x}, \hat{y})$ we can apply Lemma 5, which gives that the deterministic policy f_*^∞ with $f_*(1) = f_*(2) = f_*(3) = 1$ is an optimal solution.

Remarks

1. *Discounted MDPs with additional constraints*
   These problems have always a stationary optimal policy. The analysis for this kind of problems is much easier than for MDPs with the average reward as optimality criterion (see Kallenberg 2010).
2. *Multiple objectives*
   Some problems may have several kinds of rewards or costs, which cannot be optimized simultaneously. Assume that we want to maximize some utility for an m-tuple of immediate rewards, say utilities $u^k(R)$ and immediate rewards $r_i^k(a)$, $(i, a) \in S \times A$, for $k = 1, 2, \ldots, m$. For each k one can find an optimal policy R_k, i.e., $u_i^k(R_k) \geq u_i^k(R)$, $i \in S$, for all policies R. However, in general, $R_k \neq R_l$ if $k \neq l$, and there does not exist one policy which is optimal for all m rewards simultaneously for all starting states. Therefore, we consider the utility function with respect a given initial distribution β. Given this initial distribution β and a policy R, we denote the utilities by $u^k(\beta, R)$. The goal in multi-objective optimization is to find an *β-efficient solution*, i.e., a policy R_* such that there exists *no other policy* R satisfying $u^k(\beta, R) \geq u^k(\beta, R_*)$ for all k and $u^k(\beta, R) > u^k(\beta, R_*)$ for at least one k. These problems can be solved, for both discounted rewards and average rewards, by CMDPs (for more details, see Kallenberg 2010).

5 Applications

5.1 Optimal stopping problems

In Chap. 8 of Derman's book (Derman 1970) optimal stopping of a Markov chain is discussed. Derman considers the following model. Let $\{X_t, t = 1, 2, \dots\}$ be a finite Markov chain with state space S and stationary transition probabilities p_{ij}. Let us suppose there exists an absorbing state 0, i.e., $p_{00} = 1$, such that $\mathbb{P}\{X_t = 0 \text{ for some } t \geq 1 \mid X_1 = i\} = 1$ for every $i \in S$. Let r_i, $i \in S$, denote nonnegative values.

When the chain is absorbed at state 0, we can think of the process as having been stopped at that point in time and we receive the value r_0. However, we can also think of stopping the process at any point in time prior to absorption and receiving the value r_i if i is the state of the chain when the process is stopped. If our aim is to receive the highest possible value and if $r_0 < \max_{i \in S} r_i$, then clearly we would not necessarily wait for absorption before stopping the process.

By a *stopping time* τ, we mean a rule that prescribes the time to stop the process. Optimal stopping of a Markov chain is the problem to determine the stopping time τ such that $\mathbb{E}\{r_{X_\tau} \mid X_1 = i\}$ is maximized for all $i \in S$. Let $M_i = \max_\tau \mathbb{E}\{r_{X_\tau} \mid X_1 = i\}$, $i \in S$. Derman has shown the following result.

Theorem 10 *If v^* is an optimal solution of the linear program*

$$\min \left\{ \sum_j v_j \;\middle|\; \begin{array}{ll} v_i \geq r_i, & i \in S \\ v_i \geq \sum_j p_{ij} v_j, & i \in S \end{array} \right\}, \tag{23}$$

then $M_i = v_i^*$, $i \in S$.

In Kallenberg (1983) this approach is generalized in the following way:

- the assumption $r_i \geq 0$, $i \in S$, is omitted;
- if we continue in state i, a cost c_i is incurred for all $i \in S$;
- we can determine not only M_i, $i \in S$, but also the states S_0 in which it is optimal to stop.

The results are based on properties for convergent MDPs with as optimality criterion the *total expected reward over an infinite horizon*. The following theorem shows the result.

Theorem 11 *Let v^* and (x^*, y^*) be optimal solutions of the following dual pair of linear programs*

$$\min \left\{ \sum_j v_j \;\middle|\; \begin{array}{ll} v_i \geq r_i, & i \in S \\ v_i \geq -c_i + \sum_j p_{ij} v_j, & i \in S \end{array} \right\} \tag{24}$$

and

$$\max \left\{ \sum_i r_i x_i - \sum_i c_i y_i \;\middle|\; \begin{array}{ll} x_j + y_j - \sum_i p_{ij} y_i = 1, & j \in S \\ x_i, y_i \geq 0, & i \in S \end{array} \right\}. \tag{25}$$

Then, $M_i = v_i^*$, $i \in S$ and $S_0 = \{i \in S \mid x_i^* > 0\}$.

Furthermore, we have the following result for *monotone* optimal stopping problems, i.e., problems that satisfy $p_{ij} = 0$ for all $i \in S_1$, $j \notin S_1$, where $S_1 = \{i \in S \mid r_i \geq -c_i + \sum_j p_{ij} r_j\}$. So, S_1 is the set of states in which immediate stopping is not worse than continuing for one period and than choose to stop. The set S_1 follows directly from the data of the model.

Theorem 12 *In a monotone optimal stopping problem a one-step look ahead policy, i.e., a policy that stops in the states of S_1 and continues outside S_1, is an optimal policy.*

5.2 Replacement problems

5.2.1 General replacement problem

In a general replacement model we have state space $S = \{0, 1, \ldots, N\}$, where state 0 corresponds to a new item, and action sets $A(0) = \{1\}$ and $A(i) = \{0, 1\}$, $i \neq 0$, where action 0 means replacing the 'old' item by a new item. We consider in this model costs instead of rewards. Let c be the cost of a new item.

Furthermore, assume that an item of state i has trade-in-value s_i and maintenance costs c_i. If in state i action 0 is chosen, then $c_i(0) = c - s_i + c_0$ and $p_{ij}(0) = p_{0j}$, $j \in S$; for action 1, we have $c_i(1) = c_i$ and $p_{ij}(1) = p_{ij}$, $j \in S$. In contrast with other replacement models, where the state is determined by the age of the item, we allow that the state of the item may change to any other state.

In this case the optimal replacement policy is in general not a control-limit rule. As optimality criterion we consider the discounted reward. For this model the primal linear program is:

$$\min \left\{ \sum_{j=0}^{N} \beta_j v_j \;\middle|\; \begin{array}{ll} \sum_{j=0}^{N} (\delta_{ij} - \alpha p_{0j}) v_j \geq -c + s_i - c_0, & 1 \leq i \leq N \\ \sum_{j=0}^{N} (\delta_{ij} - \alpha p_{ij}) v_j \geq -c_i, & 0 \leq i \leq N \end{array} \right\}, \tag{26}$$

where $\beta_j > 0$, $j \in S$. Because there is only one action in state 0, namely action 1, we have $v_0^\alpha = -c_0 + \alpha \sum_{j=0}^{N} p_{0j} v_j^\alpha$.

Hence, instead of $v_i - \alpha \sum_{j=0}^{N} p_{0j} v_j = \sum_{j=0}^{N} (\delta_{ij} - \alpha p_{0j}) v_j \geq -c + s_i - c_0$, we can write $v_i - v_0 \geq -c + s_i$, obtaining the equivalent linear program

$$\min \left\{ \sum_{j=0}^{N} \beta_j v_j \;\middle|\; \begin{array}{ll} v_i - v_0 \geq r_i, & 1 \leq i \leq N \\ \sum_{j=0}^{N} (\delta_{ij} - \alpha p_{ij}) v_j \geq -c_i, & 0 \leq i \leq N \end{array} \right\}, \tag{27}$$

where $r_i = -c + s_i$, $i \in S$. The dual linear program of (27) is:

$$\max \left\{ \sum_{i=1}^{N} r_i x_i - \sum_{i=0}^{N} c_i y_i \;\middle|\; \begin{array}{ll} -\sum_{i=1}^{N} x_i + \sum_{i=0}^{N} (\delta_{i0} - \alpha p_{i0}) y_i = \beta_0 & \\ x_j + \sum_{i=0}^{N} (\delta_{ij} - \alpha p_{ij}) y_i = \beta_j, & 1 \leq j \leq N \\ x_i \geq 0, & 1 \leq i \leq N \\ y_i \geq 0, & 0 \leq i \leq N \end{array} \right\}. \tag{28}$$

For this linear program the following result can be shown. For the proof we refer to Kallenberg (2010).

Theorem 13 *There is a one-to-one correspondence between the extreme solutions of (28) and the set of deterministic policies.*

Consider the simplex method to solve (28) and start with the basic solution that corresponds to the policy which chooses action 1 (no replacement) in all states. Hence, in the first simplex tableau y_j, $0 \le j \le N$, are the basic variables and x_i, $1 \le i \le N$, the nonbasic variables. Take the usual version of the simplex method in which the column with the most negative cost is chosen as pivot column. It turns out, see Theorem 14, that this choice gives the optimal action for that state, i.e., in that state action 0, the replacement action, is optimal. Hence, after interchanging x_i and y_i, the column of y_i can be deleted. Consequently, we obtain the following *greedy simplex algorithm*.

Algorithm 3 (Greedy simplex algorithm)

1. Start with the basic solution corresponding to the nonreplacing actions.
2. If the reduced costs are nonnegative: the corresponding policy is optimal (STOP).
   Otherwise:
   (a) Choose the column with the most negative reduced cost as pivot column.
   (b) Execute the usual simplex transformation and delete the pivot column.
3. If all columns are removed: replacement in all states is the optimal policy (STOP).
   Otherwise: return to step 2.

Theorem 14 *The greedy simplex algorithm is correct and has complexity* $\mathcal{O}(N^3)$.

Remark 1 For the proof of Theorem 14 we also refer to Kallenberg (2010). The linear programming approach, as discussed in this section, is related to a paper by Gal (1984), in which the method of policy iteration was considered.

Remark 2 An optimal stopping problem may be considered as a special case of a replacement problem with as optimality criterion the total expected reward, i.e., $\alpha = 1$. In an optimal stopping problem there are two actions in each state. The first action is the stopping action and the second action corresponds to continue. If the stopping action is chosen in state i, then a final reward r_i is earned and the process terminates. If the second action is chosen, then a cost c_i is incurred and the transition probability of being in state j at the next decision time point is p_{ij}, $j \in S$. This optimal stopping problem is a special case of the replacement problem with $p_{0j} = 0$ for all $j \in S$, $c_i(0) = -r_i$ and $c_i(1) = c_i$ for all $i \in S$. Hence, also for the optimal stopping problem, the linear programming approach of this section can be used and the complexity is also $\mathcal{O}(N^3)$.

Remark 3 With a similar approach, the average reward criterion for an irreducible general replacement problem can be treated.

5.2.2 Replacement problem with increasing deterioration

Consider a replacement model with state space $S = \{0, 1, \ldots, N+1\}$. An item is in state 0 if and only if it is new; an item is in state $N+1$ if and only if it is inoperative. In states $1, 2, \ldots, N$ there are two actions: action 0 is to replace the item by a new one and action 1 is not to replace the item. In the states 0 and $N+1$ only one action is possible (no replacement and replacement by a new item, respectively) and call this action 1 and 0, respectively. The transition probabilities are:

$$p_{ij}(0) = \begin{cases} 0, & 1 \le i \le N+1,\ j \ne 0 \\ 1, & 1 \le i \le N+1,\ j = 0 \end{cases}; \quad p_{ij}(1) = p_{ij},\ 0 \le i \le N,\ 1 \le j \le N+1.$$

We assume two types of cost, the cost $c_0 \geq 0$ to replace an operative item by a new one and the cost $c_0 + c_1$, where $c_1 \geq 0$, to replace an inoperative item by a new one. Thus, c_1 is the additional cost incurred if the item becomes inoperative before being replaced. Hence, the costs c are:

$$c_i(0) = c_0, \quad 1 \leq i \leq N; \quad c_{N+1}(0) = c_0 + c_1; \quad c_i(1) = 0, \quad 0 \leq i \leq N.$$

We state the following assumptions, which turn out to be equivalent (see Lemma 6).

Assumption 1 The transition probabilities are such that for every nondecreasing function x_j, $j \in S$, the function $F(i) = \sum_{j=0}^{N+1} p_{ij} x_j$ is nondecreasing in i.

Assumption 2 The transition probabilities are such that for every $k \in S$, the function $G_k(i) = \sum_{j=k}^{N+1} p_{ij}$ is nondecreasing in i.

Lemma 6 *The Assumptions 1 and 2 are equivalent.*

The significance of Lemma 6 is that Assumption 1 can be verified by the verification of Assumption 2, which can be verified only using the data of the model. Assumption 2 means that this replacement model has increasing deterioration.

We first consider the criterion of discounted costs. For this criterion the following result can be shown, which is based on the property that the value vector v_i^α, $0 \leq i \leq N+1$, is nondecreasing in the states i.

Theorem 15 *If Assumption 1 (or 2) holds and if the state i_* is such that $i_* = \max\{i \mid \alpha \sum_j p_{ij} v_j^\alpha \leq c_0 + \alpha \sum_j p_{0j} v_j^\alpha\}$. Then, the control-limit policy f_*^∞ which replaces in the states $i > i_*$ is a discounted optimal policy.*

Theorem 15 implies that the next algorithm computes an optimal control-limit policy for this model. Similar to Algorithm 3 it can be shown that the complexity of Algorithm 4 is $\mathcal{O}(N^3)$.

Algorithm 4 (Computation of an optimal control-limit policy)

1. (a) Start with the basic solution corresponding to the nonreplacing actions in the states $i = 1, 2, \ldots, N$ and to the only action in the states 0 and $N+1$.
   (b) Let $k = N$ (the number of nonbasic variables corresponding to the replacing actions in the states $i = 1, 2, \ldots, N$).
2. If the reduced costs are nonnegative: the corresponding policy is optimal (STOP).
   Otherwise:
   (a) Choose the column corresponding to state k as pivot column.
   (b) Execute the usual simplex transformation.
   (c) Delete the pivot column.
3. If all columns are removed: replacement in all states is the optimal policy (STOP).
   Otherwise: return to step 2.

Next, we consider the criterion of average cost. By Theorem 15, for each $\alpha \in (0, 1)$ there exists a control-limit policy f_α^∞ that is α-discounted optimal. Let $\{\alpha_k, k = 1, 2, \ldots\}$ be any sequence of discount factors such that $\lim_{k\to\infty} \alpha_k = 1$.

Since there are only a finite number of different control-limit policies, there is a subsequence with one of these policies. Therefore, we may assume that $f_{\alpha_k}^{\infty} = f_0^{\infty}$ for all k. Let f^{∞} be any policy in $C(D)$. Since $f_0^{\infty} = f_{\alpha_k}^{\infty}$ is optimal for all k, we have

$$(1-\alpha_k)v^{\alpha_k}(f^{\infty}) \geq (1-\alpha_k)v^{\alpha_k}(f_0^{\infty}) \quad \text{for } k = 1, 2, \ldots.$$

Letting $k \to \infty$, we obtain for every $f^{\infty} \in C(D)$,

$$\phi(f^{\infty}) = \lim_{k\to\infty}(1-\alpha_k)v^{\alpha_k}(f^{\infty}) \geq \lim_{k\to\infty}(1-\alpha_k)v^{\alpha_k}(f_0^{\infty}) = \phi(f_0^{\infty}).$$

Therefore, the following result holds.

Theorem 16 *If Assumption* 1 (*or* 2) *holds, then there exists a control-limit policy* f_*^{∞} *such that* $\phi(f_*^{\infty}) \leq \phi(f^{\infty})$ *for all policies* $f^{\infty} \in C(D)$.

Remark The results of this section, with the exception of Algorithm 4, have been developed by Derman (1963).

5.2.3 Skip to the right model with failure

This model is slightly different from the previous one, replacement with increasing deterioration. Let the state space $S = \{0, 1, \ldots, N+1\}$, where state 0 corresponds to a new item and state $N+1$ to failure. The states i, $0 \leq i \leq N$, may be interpreted as the age of the item. The system has in state i $(0 \leq i \leq N)$ a failure probability p_i during the next period. When failure occurs in state i, which is modeled as being transferred to state $N+1$, there is an additional cost f_i. In state $N+1$ the item has to be replaced by a new one. In the states $1 \leq i \leq N$ there are two actions. Action 0 replaces the item immediately by a new one, so it has the same transitions as state 0; the replacement cost is c. By action 1 the system moves, when there is no failure, from state i to the next state $i+1$: the system skips to the right, i.e., the age of the item increases. Furthermore, in state i there are maintenance cost c_i.

The action sets, the cost of a new item, the maintenance costs and the transition probabilities are as follows.

$$A(0) = \{1\}; \quad A(i) = \{0, 1\}, \quad 1 \leq i \leq N; \quad A(N+1) = \{0\}.$$

$$1 \leq i \leq N+1 : p_{ij}(0) = \begin{cases} 1-p_0 & j = 1 \\ p_0 & j = N+1 \end{cases}; \quad c_i(0) = c + c_0 + p_0 f_0$$

$$0 \leq i \leq N : p_{ij}(1) = \begin{cases} 1-p_i & j = i+1 \\ p_i & j = N+1 \end{cases}; \quad c_i(1) = c_i + p_i f_i$$

We impose the following assumptions:

(A1) $c \geq 0$; $c_i \geq 0$, $f_i \geq 0$, $0 \leq i \leq N$.
(A2) $p_0 \leq p_1 \leq \cdots \leq p_N$, i.e., older items have greater failure probability.
(A3) $c_0 + p_0 f_0 \leq c_1 + p_1 f_1 \leq \cdots \leq c_N + p_N f_N$, i.e., the expected maintenance and failure costs grow with the age of the item.

Take any $k \in S$. Since

$$\sum_{j=k}^{N+1} p_{ij}(1) = \begin{cases} p_i & i \leq k-2 \\ 1 & i \geq k-1 \end{cases},$$

this summation is, by assumption A2, nondeceasing in i. Hence, Assumption 2 and consequently also Assumption 1 of the previous section, is satisfied. This enables us to treat this model in a similar way as the model with increasing deterioration. In this way we can derive the following result.

Theorem 17 *Let the assumptions* (A1), (A2) *and* (A3) *hold, and let* $i_* = \max\{i \mid c_i + p_i f_i + \alpha \sum_j p_{ij}(1) v_j^\alpha \leq c + c_0 + p_0 f_0 + \alpha \sum_j p_{0j}(1) v_j^\alpha\}$. *Then, the control-limit policy* f_*^∞ *which replaces in the states* $i > i_*$ *is an optimal policy.*

Remarks

1. For the proof of Theorem 17 we refer to Kallenberg (1994).
2. Algorithm 4 is also applicable to this model.
3. Similarly as in the previous section it can be shown that for the average cost criterion there exists also a control-limit optimal policy.
4. In Derman (1970, pp. 125–130) a surveillance-maintenance-replacement model is discussed. This model is solved in the following way:
   (a) A fractional linear programming formulation is developed from which an optimal policy can be derived.
   (b) This fractional linear programming can be transformed into a normal linear program. This transformation is due to Derman and his student Klein (see Derman 1962 and Klein 1962). See Charnes and Cooper (1962) and Wagner and Yuan (1968) for more general treatment of linear fractional programming.

5.2.4 Separable replacement problem

Suppose that the MDP has the following structure: $S = \{0, 1, 2, \ldots, N\}$; $A(i) = \{1, 2, \ldots, M\}$, $i \in S$; $p_{ij}(a) = p_j(a)$, $i, j \in S$, $a \in A(i)$, i.e., the transitions are state independent; $r_i(a) = s_i + t(a)$, $i \in S$, $a \in A(i)$, i.e., the rewards are separable.

As example, consider the problem of periodically replacing a car. The age of a car can be $0, 1, \ldots, N$. When a car is replaced, it can be replaced not only by a new one (state 0), but also by a car in an arbitrary state a, $1 \leq a \leq N$. Let s_i be the trade-in-value of a car of state i, $t(a)$ the costs of a car of state a. Then, $r_i(a) = s_i - t(a)$ and $p_{ij}(a) = p_j(a)$, where $p_j(a)$ is the probability that a car of state a is in state j at the next decision time point.

The next theorems show that a one-step look ahead policy is optimal both for discounted as for undiscounted rewards.

Theorem 18 *The policy* f_1^∞, *defined by* $f_1(i) = a_1$ *for all* i, *where* a_1 *is such that* $-t(a_1) + \alpha \sum_j p_j(a_1) s_j = \max_{1 \leq a \leq M}\{-t(a) + \alpha \sum_j p_j(a) s_j\}$, *is an* α*-discounted optimal policy.*

Theorem 19 *The policy* f_2^∞, *defined by* $f_2(i) = a_2$ *for all* i, *where* a_2 *is such that* $-t(a_2) + \sum_j p_j(a_2) s_j = \max_{1 \leq a \leq M}\{-t(a) + \sum_j p_j(a) s_j\}$, *is an average optimal policy.*

5.3 Multi-armed bandit problems

5.3.1 Introduction

The multi-armed bandit problem is a model for dynamic allocation of a resource to one of n independent alternative projects. Any project may be in one of a finite number of states, say project j in the set S_j, $j = 1, 2, \ldots, n$. Hence, the state space S is the Cartesian product $S = S_1 \times S_2 \times \cdots \times S_n$. Each state $i = (i_1, i_2, \ldots, i_n)$ has the same action set $A = \{1, 2, \ldots, n\}$, where action k means that project k is chosen, $k = 1, 2, \ldots, n$. So, at each stage one can be working on exactly one of the projects.

When project k is chosen in state i—the chosen project is called the *active project*—the immediate reward and the transition probabilities only depend on the active project, whereas the states of the remaining projects are frozen. Let r_{i_k} and $p_{i_k j}$, $j \in S_k$ denote these quantities when action k is chosen. The total discounted reward criterion is chosen.

It was shown by Gittins and Jones (1974, 1979) that an optimal policy is the policy that selects project k in state $i = (i_1, i_2, \ldots, i_n)$, where k satisfies

$$G_k(i_k) = \max_{1 \le j \le n} G_j(i_j)$$

for certain numbers $G_j(i_j)$, $i_j \in S_j$, $1 \le j \le n$. Such a policy is called an *index policy*. Surprisingly, the number $G_j(i_j)$ only depends on project j and not on the other projects. These indices are called the *Gittins indices*.

As a consequence, the multi-armed bandit problem can be solved by solving a sequence of n one-armed bandit problems. This is a *decomposition* result by which the dimensionality of the problem is reduced considerably. Algorithms with complexity $\mathcal{O}(\sum_{j=1}^{n} n_j^3)$, where $n_j = |S_j|$, $1 \le j \le n$, do exist for the computation of all indices.

5.3.2 A single project with a terminal reward

Consider the one-armed bandit problem with stopping option, i.e., in each state there are two options: action 1 is the stopping option and then one earns a terminal reward M and by action 2 the process continues with in state i an immediate reward r_i and transition probabilities p_{ij}. Let $v^\alpha(M)$ be the value vector of this optimal stopping problem. Then, $v^\alpha(M)$ is the unique solution of the optimality equation

$$v_i^\alpha(M) = \max\left\{M, r_i + \alpha \sum_j p_{ij} v_j^\alpha(M)\right\}, \quad i \in S, \tag{29}$$

and of the linear program

$$\min\left\{\sum_j v_j \,\middle|\, \begin{array}{ll} \sum_j \{\delta_{ij} - \alpha p_{ij}\} v_j \ge r_i, & i \in S \\ v_i \ge M, & i \in S \end{array}\right\}. \tag{30}$$

Furthermore, we have the following results.

Theorem 20 *Let* (x, y) *be an extreme optimal solution of the dual program of* (30), *i.e.*,

$$\max\left\{\sum_j r_i x_i + M \cdot \sum_j y_i \,\middle|\, \begin{array}{ll} \sum_i \{\delta_{ij} - \alpha p_{ij}\} x_i + y_j = 1, & i \in S \\ x_i, y_i \ge 0, & i \in S \end{array}\right\}. \tag{31}$$

Then, the policy f^∞ such that

$$f(i) = \begin{cases} 2 & \text{if } x_i > 0 \\ 1 & \text{if } x_i = 0 \end{cases}$$

is an optimal policy.

Lemma 7 *$v_i^\alpha(M) - M$ is a nonnegative continuous nonincreasing function in M, for all $i \in S$.*

Define the indices G_i, $i \in S$, by $G_i = \min\{M \mid v_i^\alpha(M) = M\}$. Hence, $v_i^\alpha(G_i) = G_i$ and, by Lemma 7, $v_i^\alpha(M) = M$ for all $M \geq G_i$. For these indices one can show the following theorem.

Theorem 21 *For any M, the policy $f^\infty \in C(D)$ which chooses the stopping action in state i if and only if $M \geq G_i$ is optimal.*

For $M = G_i$ both actions (stop or continue) are optimal. Hence, an interpretation of the Gittins index G_i is that it is the terminal reward under which in state i both actions are optimal. Therefore, this number is also called the *indifference value*.

5.3.3 Multi-armed bandits

Consider the multi-armed bandit model with an additional option (action 0) in each state. Action 0 is a stopping option and then one earns a terminal reward M. One can show the following result.

Theorem 22 *For any state $i = (i_1, i_2, \ldots, i_n)$ and any terminal reward M, the policy that takes the stopping action if $M \geq G_{i_j}$ for all $j = 1, 2, \ldots, n$ and continues with project k if $G_{i_k} = \max_j G_{i_j} > M$, is an optimal policy.*

The preceding theorem shows that the optimal policy in the multi-project case can be determined by an analysis of the n single-project problems, with the optimal decision in state $i = (i_1, i_2, \ldots, i_n)$ being to operate on that project k having the largest G_{i_k} if this value is greater than M and to stop otherwise.

Several methods have been proposed for the computation of the Gittins indices. We mention the contributions of Katehakis and Veinott (the restart-in-state method, see Katehakis and Veinott 1987), Varaiya, Walrand and Buyukkoc (the-largest-remaining-index method, see Varaiya et al. 1985), and Chen and Katehakis (the linear programming method, see Chen and Katehakis 1986). In this article we present the parametric linear programming method proposed in Kallenberg (1986). This method has for a project with N states complexity $\mathcal{O}(N^3)$.

5.3.4 The parametric linear programming method

We have already seen that for a single project with terminal reward M the solution can be obtained from a linear programming problem, namely program (31). For M big enough, e.g., for $M \geq C = (1-\alpha) \cdot \max_i r_i$, we know that $v_i^\alpha(M) = M$ for all states i. Furthermore, we have seen that the Gittins index $G_i = \min\{M \mid v_i^\alpha(M) = M\}$.

One can solve program (31) as a parametric linear programming problem with parameter M. Starting with $M = C$ one can decrease M and find for each state i the largest M for which it is optimal to keep working on the project, which is in fact $\min\{M \mid v_i^{\alpha}(M) = M\} = G_i$, in the order of decreasing M-values.

One can start with the simplex tableau in which all y-variables are in the basis and in which the x-variables are the nonbasic variables. This tableau is optimal for $M \geq C$. Decrease M until we meet a basis change, say the basic variable y_i will be exchanged with the nonbasic variable x_i. Then, we know the M-value which is equal to G_i. In this way we continue and repeat the procedure N times, where N is the number of states in the current project. The used pivoting row and column do not influence any further pivoting step, so we can delete these row and column from the simplex tableau.

We can easily determine the computational complexity. Each update of an element in a simplex tableau needs at most two arithmetic operations (multiplication and divisions as well as additions and subtractions). Hence, the total number of arithmetic operations in this method for a project with N states, is at most $2 \cdot \sum_{k=1}^{N} k^2 = \frac{1}{3}N(N+1)(2N+1) = \mathcal{O}(N^3)$.

Remark The problem of assigning one of several treatments in clinical trials can be formulated as a multi-armed bandit problem. Derman and Katehakis (1987) have used the characterization of the Gittins index as a restart-in-state problem (see Katehakis and Veinott 1987) to calculate efficiently the Gittins values for clinical trials. The characterization of the Gittins index as a restart-in-state problem is related to a general replacement problem as treated by Derman in his book (Derman 1970, pp. 121–125).

5.4 Separable problems

5.4.1 Introduction

Separable MDPs have the property that for certain pairs $(i, a) \in S \times A$:

(1) the immediate reward is the sum of two terms, one depends only on the current state and the other depends only on the chosen action: $r_i(a) = s_i + t_a$.
(2) the transition probabilities depend only on the action and not on the state from which the transition occurs: $p_{ij}(a) = p_j(a)$, $j \in S$.

Let $S_1 \times A_1$ be the subset of $S \times A$ for which the pairs (i, a) satisfy (1) and (2). We also assume that the action sets of A_1 are *nested*: let $S_1 = \{1, 2, \ldots, m\}$, then $A_1(1) \supseteq A_1(2) \supseteq \cdots \supseteq A_1(m) \neq \emptyset$. Let $S_2 = S \backslash S_1$, $A_2(i) = A(i) \backslash A_1(i)$, $1 \leq i \leq m$ and $A_2(i) = A(i)$, $m + 1 \leq i \leq N$. We also introduce the notation $B(i) = A_1(i) \backslash A_1(i+1)$, $1 \leq i \leq m-1$ and $B(m) = A_1(m)$. Then, $A_1(i) = \bigcup_{j=i}^{m} B(j)$ and the sets $B(j)$ are disjunct. We allow that S_2, A_2 or $B(i)$, $1 \leq i \leq m-1$, are empty sets.

If the system is observed in state $i \in S_1$ and the decision maker will choose an action from $A_1(i)$, then, the decision process can be considered as follows. First, a reward s_i is earned and the system makes a zero-time transition to an additional state $N + i$. In this additional state there are two options: either to take an action $a \in B(i)$ or to take an action $a \in A_1(i) \backslash B(i) = A_1(i+1)$. In the first case the reward t_a is earned and the process moves to state j with probability $p_j(a)$, $j \in S$; in the second case we are in the same situation as in state $N + i$, but now in $N + i + 1$, i.e., a zero-time transition is made from state $N + i$ to state $N + i + 1$.

A lot of dynamic decision problems are separable, e.g., the automobile replacement problem which was first considered by Howard (see Howard 1960)

5.4.2 Discounted rewards

The description in the introduction as a problem with zero-time and one-time transitions gives rise to the transformed model with $N+m$ states and to the following linear program for the computation of the value vector v^α:

$$\min\left\{\sum_{i=1}^{N} v_i + \sum_{i=1}^{m} y_i \;\middle|\; \begin{array}{ll} v_i \geq r_i(a) + \alpha \sum_{j=1}^{N} p_{ij}(a) v_j, & 1 \leq i \leq N,\ a \in A_2(i) \\ v_i \geq s_i + y_i, & 1 \leq i \leq m \\ y_i \geq t_a + \alpha \sum_{j=1}^{N} p_j(a) v_j, & 1 \leq i \leq m,\ a \in B(i) \\ y_i \geq y_{i+1}, & 1 \leq i \leq m-1 \end{array}\right\}. \tag{32}$$

The first set of inequalities corresponds to the non-separable set $S \times A_2$ with one-time transitions; the second set inequalities to the zero-time transitions from the state i to $N+i$, $1 \leq i \leq m$; the third set of inequalities to the set $S_1 \times B$ with one-time transitions and the last set inequalities corresponds to the zero-time transitions from the state $N+i$ to $N+i+1$, $1 \leq i \leq m-1$.

The dual of program (32), where the dual variables $x_i(a)$, λ_i, $w_i(a)$, ρ_i correspond to the four sets of constraints in (32), is:

$$\max \sum_{i=1}^{N} \sum_{a \in A_2(i)} r_i(a) x_i(a) + \sum_{i=1}^{m} s_i \lambda_i + \sum_{i=1}^{m} \sum_{a \in B(i)} w_i(a) \tag{33}$$

subject to the constraints

$$\sum_{i=1}^{N} \sum_{a \in A_2(i)} \{\delta_{ij} - \alpha p_{ij}(a)\} x_i(a) + \sum_{i=1}^{m} \delta_{ij} \lambda_i - \sum_{i=1}^{m} \sum_{a \in B(i)} p_j(a) w_i(a) = 1, \quad 1 \leq j \leq N$$

$$\rho_j - \rho_{j-1} - \lambda_j + \sum_{a \in B(j)} w_j(a) = 1, \quad 1 \leq j \leq m-1$$

$$-\rho_{m-1} - \lambda_m + \sum_{a \in B(m)} w_m(a) = 1$$

$$x_i(a) \geq 0,\ 1 \leq i \leq N,\ a \in A_2(i);\ \lambda_i \geq 0,\ 1 \leq i \leq m;$$

$$w_i(a) \geq 0,\ 1 \leq i \leq m,\ a \in B(i);\ \rho_i \geq 0,\ 1 \leq i \leq m-1.$$

Without using the transformed problem, the linear program to compute the value vector v^α is:

$$\min\left\{\sum_{i=1}^{N} v_i \;\middle|\; v_i \geq r_i(a) + \alpha \sum_{j=1}^{N} p_{ij}(a) v_j,\ 1 \leq i \leq N,\ a \in A(i)\right\}. \tag{34}$$

The following result can be shown.

Lemma 8 *Let the vector v be feasible for (34) and define the vector y by $y_i = \max_{a \in A_1(i)}\{t_a + \alpha \sum_{j=1}^{N} p_j(a) v_j\}$, $1 \leq i \leq m$. Then,*

(1) *(v, y) is a feasible solution of (32).*

(2) $\sum_{i=1}^{N} v_i + \sum_{i=1}^{m} y_i \geq \sum_{i=1}^{N} v_i^{\alpha} + \sum_{i=1}^{m} \max_{a \in A_1(i)} \{t_a + \alpha \sum_{j=1}^{N} p_j(a) v_j^{\alpha}\}$.

Since v^{α} is the unique optimal solution of (34), we have shown that (v^{α}, y^{α}), with $y_i^{\alpha} = \max_{a \in A_1(i)} \{t_a + \alpha \sum_{j=1}^{N} p_j(a) v_j^{\alpha}\}$, $1 \leq i \leq m$, is the unique optimal solution of (32). The next theorem shows how an optimal policy can be found from an optimal solution of problem (33).

Theorem 23 *Let* $(x^*, \lambda^*, w^*, \rho^*)$ *be an optimal solution of* (33). *Define* $S_* = \{j \mid \sum_{a \in A_2(j)} x_j^*(a) > 0\}$ *and* $k_j = \min\{k \geq j \mid \sum_{a \in B(k)} w_k^*(a) > 0\}$, $j \in S \backslash S_*$. *Take any policy* $f_*^{\infty} \in C(D)$ *such that* $x_j^*(f_*(j)) > 0$ *if* $j \in S_*$ *and* $w_{k_j}^*(f_*(j)) > 0$ *if* $j \in S \backslash S_*$. *Then,* f_*^{∞} *is well-defined and a discounted optimal policy.*

5.4.3 Average rewards—unichain case

Consider the problem again in the transformed model with $N + m$ states and with zero-time and one-time transitions. This interpretation gives rise to the following linear program for the computation of the value vector ϕ.

$$\min \left\{ x \;\middle|\; \begin{array}{ll} x + y_i \geq r_i(a) + \sum_{j=1}^{N} p_{ij}(a) y_j, & 1 \leq i \leq N,\ a \in A_2(i) \\ y_i \geq s_i + z_i, & 1 \leq i \leq m \\ x + z_i \geq t_a + \sum_{j=1}^{N} p_j(a) y_j, & 1 \leq i \leq m,\ a \in B(i) \\ z_i \geq z_{i+1}, & 1 \leq i \leq m-1 \end{array} \right\}. \tag{35}$$

The dual of program (35), where the dual variables $x_i(a)$, λ_i, $w_i(a)$, ρ_i correspond to the four sets of constraints in (35), is:

$$\max \sum_{i=1}^{N} \sum_{a \in A_2(i)} r_i(a) x_i(a) + \sum_{i=1}^{m} s_i \lambda_i + \sum_{i=1}^{m} \sum_{a \in B(i)} w_i(a) \tag{36}$$

subject to the constraints

$$\sum_{i=1}^{N} \sum_{a \in A_2(i)} \{\delta_{ij} - p_{ij}(a)\} x_i(a) + \sum_{i=1}^{m} \delta_{ij} \lambda_i - \sum_{i=1}^{m} \sum_{a \in B(i)} p_j(a) w_i(a) = 0, \quad 1 \leq j \leq N$$

$$\rho_j - \rho_{j-1} - \lambda_j + \sum_{a \in B(j)} w_j(a) = 0, \quad 1 \leq j \leq m-1$$

$$-\rho_{m-1} - \lambda_m + \sum_{a \in B(m)} w_m(a) = 0$$

$$\sum_{i=1}^{N} \sum_{a \in A_2(i)} x_i(a) + \sum_{i=1}^{m} \sum_{a \in B(i)} w_i(a) = 1$$

$x_i(a) \geq 0,\ 1 \leq i \leq N,\ a \in A_2(i);\ \lambda_i \geq 0,\ 1 \leq i \leq m;$

$w_i(a) \geq 0,\ 1 \leq i \leq m,\ a \in B(i);\ \rho_0 = 0;\ \rho_i \geq 0,\ 1 \leq i \leq m-1.$

Without using the transformed problem, the linear program to compute the value ϕ is:

$$\min\left\{x \,\middle|\, x + y_i \geq r_i(a) + \sum_{j=1}^{N} p_{ij}(a)y_j,\ 1 \leq i \leq N,\ a \in A(i)\right\}. \tag{37}$$

Lemma 9 *Let (x, y) feasible for problem (37) and define the vector z by $z_i = \max_{a\in A_1(i)}\{t_a + \sum_{j=1}^{N} p_j(a)y_j\} - x$, $1 \leq i \leq m$. Then, (x, y, z) is a feasible solution of (35) and $x \geq \phi$.*

Since any optimal solution (x^*, y^*) of problem (37) satisfies $x^* = \phi$, the optimum value of (35) is also ϕ. Furthermore, $(x^* = \phi, y^*, z^*)$ is an optimal solution of program (35), where $z_i^* = \max_{a\in A_1(i)}\{t_a + \sum_{j=1}^{N} p_j(a)y_j^*\} - \phi$ for $i = 1, 2, \ldots, m$. The next theorem shows how an optimal policy can be found from an optimal solution of problem (36).

Theorem 24 *Let $(x^*, \lambda^*, w^*, \rho^*)$ be an optimal solution of (36). Define $S_* = \{j \mid \sum_{a\in A_2(j)} x_j^*(a) > 0\}$ and $k_j = \min\{k \geq j \mid \sum_{a\in B(k)} w_k^*(a) > 0\}$, $j \in S_{w^*}$, where $S_{w^*} = \{j \in S\backslash S_* \mid \sum_{a\in A_1(j)} w_j^*(a) > 0\}$. Take any policy $f_*^\infty \in C(D)$ such that $x_j^*(f_*(j)) > 0$ if $j \in S_*$, $w_{k_j}^*(f_*(j)) > 0$ if $j \in S_{w^*}$ and $f_*(j)$ arbitrarily chosen if $j \notin S_* \cup S_{w^*}$. Then, f_*^∞ is an average optimal policy.*

5.4.4 Average rewards—general case

Again, the interpretation of the transformed model gives rise to consider the following linear program in order to compute the value vector ϕ.

$$\min\left\{\sum_{j=1}^{N} x_j + \sum_{j=1}^{m} w_j \,\middle|\, \begin{array}{ll} x_i \geq \sum_{j=1}^{N} p_{ij}(a)x_j, & 1 \leq i \leq N,\ a \in A_2(i) \\ x_i \geq w_i, & 1 \leq i \leq m \\ w_i \geq \sum_{j=1}^{N} p_j(a)x_j, & 1 \leq i \leq m,\ a \in B(i) \\ w_i \geq w_{i+1}, & 1 \leq i \leq m-1 \\ x_i + y_i \geq r_i(a) + \sum_{j=1}^{N} p_{ij}(a)y_j, & 1 \leq i \leq N,\ a \in A_2(i) \\ y_i \geq s_i + z_i, & 1 \leq i \leq m \\ w_i + z_i \geq t_a + \sum_{j=1}^{N} p_j(a)y_j, & 1 \leq i \leq m,\ a \in B(i) \\ z_i \geq z_{i+1}, & 1 \leq i \leq m-1 \end{array}\right\}. \tag{38}$$

The dual of program (38), where the dual variables $y_i(a)$, μ_i, $z_i(a)$, σ_i, $x_i(a)$, λ_i, $w_i(a)$, ρ_i correspond to the eight sets of constraints in (38), is:

$$\max \sum_{i=1}^{N} \sum_{a\in A_2(i)} r_i(a)x_i(a) + \sum_{i=1}^{m} s_i\lambda_i + \sum_{i=1}^{m} \sum_{a\in B(i)} t_a w_i(a) \tag{39}$$

subject to the constraints

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{a\in A_2(i)}\{\delta_{ij}-p_{ij}(a)\}y_i(a)+\sum_{i=1}^{m}\delta_{ij}\mu_i \\
&\qquad -\sum_{i=1}^{m}\sum_{a\in B(i)}p_j(a)z_i(a)+\sum_{a\in A_2(i)}x_j(a) = 1,\ 1\le j\le N \\
&\sigma_j-\sigma_{j-1}-\mu_j+\sum_{a\in B(j)}w_j(a)+\sum_{a\in B(j)}z_j(a) = 1,\ 1\le j\le m \\
&\sum_{i=1}^{N}\sum_{a\in A_2(i)}\{\delta_{ij}-p_{ij}(a)\}x_i(a)+\sum_{i=1}^{m}\delta_{ij}\lambda_i \\
&\qquad -\sum_{i=1}^{m}\sum_{a\in B(i)}p_j(a)w_i(a) = 0,\ 1\le j\le N \\
&\rho_j-\rho_{j-1}-\lambda_j+\sum_{a\in B(j)}w_j(a) = 0,\ 1\le j\le m \\
&\rho_0=\rho_m=\sigma_0=\sigma_m=0;\ x_i(a),\ y_i(a),\ z_i(a),\ w_i(a),\ \lambda_i,\ \mu_i,\ \rho_i,\ \sigma_i\ge 0
\end{aligned}$$

for all i and a.

Without using the transformed problem, the linear program to compute the value ϕ is:

$$\min\left\{\sum_{j=1}^{N}x_j\ \middle|\ \begin{array}{ll}\sum_{j=1}^{N}\{\delta_{ij}-p_{ij}(a)\}x_j\ge 0, & 1\le i\le N,\ a\in A(i)\\ x_i+\sum_{j=1}^{N}\{\delta_{ij}-p_{ij}(a)\}u_j\ge r_i(a), & 1\le i\le N,\ a\in A(i)\end{array}\right\}. \tag{40}$$

Theorem 25 *Let* (x^*,w^*,y^*,z^*) *and* $(y^*,\mu^*,z^*,\sigma^*,x^*,\lambda^*,w^*,\rho^*)$ *be optimal solutions of the problems* (38) *and* (39)*, respectively. Let* m_i *and* n_i *defined by* $m_i=\min\{j\ge i\mid \sum_{a\in B(j)}w_j^*(a)>0\}$ *and* $n_i=\min\{j\ge i\mid \sum_{a\in B(j)}\{w_j^*(a)+z_j^*(a)\}>0\}$*. Take any policy* $f_*^\infty\in C(D)$ *such that*

$$\begin{aligned}
&x_i^*\big(f_*(i)\big)>0 && \text{if } i\in S_*,\ \text{where } S_*=\textstyle\sum_{a\in A_2(i)}x_i^*(a)>0;\\
&w_{m_i}^*\big(f_*(i)\big)>0 && \text{if } i\notin S_*\ \text{and}\ \lambda_i^*>0;\\
&y_i^*\big(f_*(i)\big)>0 && \text{if } i\notin S_*,\ \lambda_i^*=0\ \text{and}\ y_i^*\big(f_*(i)\big)>0;\\
&w_{n_i}^*\big(f_*(i)\big)>0 && \text{if } i\notin S_*,\ \lambda_i^*=\textstyle\sum_{a\in A_2(i)}y_i^*(a)=0\ \text{and}\ \sum_{a\in A_1(i)}w_{n_i}^*(a)>0;\\
&z_{n_i}^*\big(f_*(i)\big)>0 && \text{if } i\notin S_*,\ \lambda_i^*=\textstyle\sum_{a\in A_2(i)}y_i^*(a)=\sum_{a\in A_1(i)}w_{n_i}^*(a)=0.
\end{aligned}$$

Then, (1) $x^*=\phi$; (2) f_*^∞ *is well-defined and an average optimal policy.*

Remark De Ghellinck and Eppen (1967) have examined separable MDPs with the discounted rewards as optimality criterion. Denardo introduced in Denardo (1968) the notion of zero-time transitions. Discounted and averaging versions (for the unichain case) are then shown to yield special linear programming formulations. In the discounted case, the linear program is identical to that of De Ghellinck and Eppen. Kallenberg (1992) has shown that for the average reward criterion also in the multichain case a special linear program can be used to solve the original problem.

References

Charnes, A., & Cooper, W. W. (1962). Programming with linear fractional functions. *Naval Research Logistics Quarterly*, *9*, 181–186.

Chen, Y. R., & Katehakis, M. N. (1986). Linear programming for finite state bandit problems. *Mathematics of Operations Research*, *11*, 180–183.

De Ghellinck, G. T. (1960). Les problèmes de décisions sequentielles. *Cahiers du Centre d'Etudes de Recherche Opérationelle*, *2*, 161–179.

De Ghellinck, G. T., & Eppen, G. D. (1967). Linear programming solutions for separable Markovian decision problems. *Management Science*, *13*, 371–394.

Denardo, E. V. (1968). Separable Markov decision problem. *Management Science*, *14*, 451–462.

Denardo, E. V. (1970). On linear programming in a Markov decision problem. *Management Science*, *16*, 281–288.

Denardo, E. V., & Fox, B. L. (1968). Multichain Markov renewal programs. *SIAM Journal on Applied Mathematics*, *16*, 468–487.

D'Epenoux, F. (1960). Sur un problème de production et de stockage dans l'aléatoire. *Revue Française de Recherche Opérationelle*, *14*, 3–16.

Derman, C. (1962). On sequential decisions and Markov chains. *Management Science*, *9*, 16–24.

Derman, C. (1970). *Finite state Markovian decision processes*. New York: Academic Press.

Derman, C. (1963). On optimal replacement rules when changes of state are Markovian. In R. Bellman (Ed.), *Mathematical optimization techniques* (pp. 201–210). Berkeley: University of California Press.

Derman, C., & Katehakis, M. N. (1987). Computing optimal sequential allocation rules in clinical trials. In J. Van Ryzin (Ed.), *I.M.S. lecture notes—monograph series: Vol. 8. Adaptive statistical procedures and related topics* (pp. 29–39).

Filar, J. A., & Schultz, T. (1988). Communicating MDPs: Equivalence and LP properties. *Operations Research Letters*, *7*, 303–307.

Gal, S. (1984). A $\mathcal{O}(N^3)$ algorithm for optimal replacement problems. *SIAM Journal of Control and Optimization*, *22*, 902–910.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistic Society Series B*, *14*, 148–177.

Gittins, J. C., & Jones, D. M. (1974). A dynamic allocation index for the sequential design of experiments. In J. Gani (Ed.), *Progress in statistics* (pp. 241–266). Amsterdam: North Holland.

Howard, R. A. (1960). *Dynamic programming and Markov processes*. Cambridge: MIT Press.

Hordijk, A., & Kallenberg, L. C. M. (1979). Linear programming and Markov decision chains. *Management Science*, *25*, 352–362.

Hordijk, A., & Kallenberg, L. C. M. (1984). Constrained undiscounted stochastic dynamic programming. *Mathematics of Operations Research*, *9*, 276–289.

Kallenberg, L. C. M. (1980). *Linear programming and finite Markovian control problem*. PhD Thesis, University of Leiden.

Kallenberg, L. C. M. (1983). *Linear programming and finite Markovian control problem*. Mathematical Centre Tract no. 148, Amsterdam.

Kallenberg, L. C. M. (1986). A note on Katehakis and Chen's computation of the Gittins index. *Mathematics of Operations Research*, *11*, 184–186.

Kallenberg, L. C. M. (1992). Separable Markov decision problems. *OR Spektrum*, *14*, 43–52.

Kallenberg, L. C. M. (1994). Survey of linear programming for standard and nonstandard Markovian control problems. Part II: Applications. *Mathematical Methods of Operations Research*, *40*, 127–143.

Kallenberg, L. C. M. (2002). Classification problems in MDPs. In Z. How, J. A. Filar, & A. Chen (Eds.), *Markov processes and controlled Markov chains* (pp. 151–165). Boston: Kluwer.

Kallenberg, L. C. M. (2010). *Markov decision processes*. Lecture notes, University of Leiden (available at http://www.math.leidenuniv.nl/~kallenberg/Lecture-notes-MDP.pdf).

Katehakis, M. N., & Veinott, A. V. Jr. (1987). The multi-armed bandit problem: decomposition and computation. *Mathematics of Operations Research*, *12*, 262–268.

Klein, M. (1962). Inspection-maintenance-replacement schedules under Markovian deterioration. *Management Science*, *9*, 25–32.

Manne, A. S. (1960). Linear programming and sequential decisions. *Management Science*, *6*, 259–267.

Tsitsiklis, J. N. (2007). NP-hardness of checking the unichain condition in average cost MDPs. *Operations Research Letters*, *35*, 319–323.

Varaiya, P. P., Walrand, J. C., & Buyukkoc, C. (1985). Extensions of the multi-armed bandit problem: the discounted case. *IEEE Transactions on Automatic Control*, *30*, 426–439.

Wagner, H. M., & Yuan, J. S. C. (1968). Algorithmic equivalence in linear fractional programming. *Management Science*, *14*, 301–306.